viu Universidad Internacional de Valencia

A Natural Language Processing Approach for assessing Quality of Life among Patients with Chronic Diseases

Titulación: Máster en Big Data y Data Science	Alumno/a: de Paúl Smith, Samuel DNI: 43477284Q	Convocatoria:
Curso Académico 2022-2023	Director/a del TFT: Vanessa Moscardó Garcia	PRIMERA (07/2023)

De:
Planeta Formación y Universidades

Contents

List of Figures 1

List of Tables 1

1 Introduction 2
1.1 Justification . . . 2
1.2 Objectives . . . 4
1.3 Document Structure . . . 5

2 Theoretical Framework 6
2.1 Chronic Diseases and Quality of Life . . . 6
2.1.1 Quality of Life Assessment Instruments (Questionnaires) . . . 9
2.1.2 SF-36 . . . 10
2.2 Natural Language Processing . . . 12
2.2.1 Data Preprocessing . . . 15
2.2.2 Feature Extraction and Modelling . . . 16
2.3 NLP in QoL assessment . . . 24

3 Methodology 26
3.0.1 SparkNLP . . . 27
3.0.2 Cloud Computing: AWS . . . 28
3.1 First Approach: Multiclass Text Classification . . . 32
3.2 Second Approach: Binary Text Classification . . . 35
3.3 Metrics . . . 35

4 Results 37
4.1 First Approach: Results of the Muliclass Classification Problem . . . 37
4.2 Second Approach: Results of the Binary Classification Problem . . . 42

5 Conclusions 48
5.1 Conclusions . . . 48
5.2 Limitations . . . 49
5.3 Future Work . . . 50

Bibliography 51

List of Figures

1.1 Leading Causes of Mortality in Spain . . . 3

2.1 SF-36 domains (Fleishman et al., 2006) . . . 10
2.2 Simulated SF-36 result . . . 11
2.3 AI, ML, DL and NLP relationship (Vajjala et al., 2020) . . . 13
2.4 Difference between Stemming and Lematization (Vajjala et al., 2020) . 16
2.5 BoW Example (1/3) (Soo, 2023) . . . 17
2.6 BoW Example (2/3) (Soo, 2023) . . . 18
2.7 BoW Example (3/3) (Soo, 2023) . . . 18
2.8 Word Embedding Example (Soo, 2023) . . . 20
2.9 Joint Embedding Arithmetic Examples (Girdhar et al., 2023) . . . 23

3.1 Testimony example . . . 26
3.2 Labelling example . . . 26
3.3 NLP Pipeline Example . . . 28
3.4 Mapping of the indicator set to the dimensions of SF-36 . . . 30
3.5 Indicator distribution by Dimension . . . 31
3.6 Frequency of Indicator Ocurrences . . . 31
3.7 Class frquency in testimony sentences . . . 32
3.8 Class frquency in testimony sentences (subsampling) . . . 33

4.1 F1 Score for each type of model . . . 37
4.2 F1 Score for the top 5 models of each type . . . 38
4.3 Proportion of embedding types in top models . . . 39
4.4 Confusion Matrix for best Multiclass Classification Model . . . 41
4.5 Example Predictions for best Multiclass Classification Model . . . 42
4.6 F1 Score for each type of model . . . 43
4.7 F1 Score for the top 5 models of each type . . . 43
4.8 Proportion of embedding types in top models . . . 44
4.9 Confusion Matrix - Binary Classificaiton . . . 46
4.10 Example Predictions - Binary Classification . . . 47

List of Tables

4.1 10 mejores modelos obtenidos de la búsqueda de hiperparámetros . 40
4.2 10 mejores modelos obtenidos de la búsqueda de hiperparámetros . 45

Abstract

Natural Language Processing stands at the front of the Artificial Intelligence revolution in the data-driven era, where advancements occur at a rapid pace, continuously pushing the boundaries of the state-of-the-art. Meanwhile, the world faces a delicate landscape in the realm of chronic diseases, with an alarming prevalence: one in every five individuals lives with at least one chronic condition.

This study sets out to explore the convergence of these two domains by providing an extensive introduction to both subjects, aiming to establish the necessary theoretical foundations for the development of a tool that assesses the well-being of patients facing these diseases. This approach focuses on identifying Quality of Life indicators in patients' testimonials.

Through various approaches, a model based on BERT embeddings is successfully developed, achieving an impressive F1 score of 0.934 on unseen data to determine whether a sentence contains a Quality of Life indicator or not. A subsequent approach yields another model, although less promising, which distinguishes the specific referents of these indicators with a modest performance.

This study lays the groundwork for the development of models that could become part of tools designed to streamline the physician-patient interaction, simplify the tasks of medical professionals, and support informed decision-making.

Keywords: NLP, Natural Language Processing, Quality of Life, Chronic Diseases, BERT, embeddings

Acknowledgments

To all the people who have supported and stood by me throughout this journey, thank you.

1. Introduction

In the current era, marked by rapid technological advancements, Artificial Intelligence is no more a concept of science fiction since it has become a tangible reality in almost every aspect of our lives. From medicine to industry, from education to entertainment, Artificial Intelligence has become omnipresent, reshaping our interactions and redefining the boundaries of what is possible. This has sparked society's profound interest while also posing ethical, social, and philosophical challenges that deserve careful consideration. In this Master's Thesis, we will delve into an application of Artificial Intelligence and Data Science in the field of Health Sciences.

In the context of chronic diseases, patients' Quality of Life plays a fundamental role in their physical, emotional, and social well-being. In the following pages, we propose the development of a tool based on Natural Language Processing techniques to identify features or indicators of patients' Quality of Life present in their testimonies. The main objective is to provide an efficient and automated system for obtaining relevant information about the subjective experience of patients with chronic diseases. This could contribute to a better understanding of their health status and enable healthcare professionals to make better informed clinical decisions.

The project we will develop in the following pages is part of an initiative that will lead to the publication of a research article. This Master's Thesis only represents the first approach of a much more complex project carried out by the ProHealth research group at the Valencia International University, of which I am currently part of.

1.1 Justification

Since the final years of my undergraduate degree in Mathematics, which I completed with honors in the corresponding Data Analysis subject, I have become passionate about Data Science. However, I have always had an interest in numerous areas of knowledge. One academic discipline that I had always regretted not exploring is the field of Health Sciences. The opportunity to work on a Master's Thesis focused on a topic like the one we are addressing presents itself as the ideal occasion to satisfy that curiosity that has always accompanied me.

In addition, I had a great interest in participating in a text-based data project, as my Bachelor's Thesis focused on time series analysis, and in my current employment, my work revolves around projects related to images and computer vision. I believe that the experience gained from working with these three major types of data provides me with the necessary foundations to become a well-rounded and versatile Data Scientist.

My primary motivation in immersing myself into Data Science is not to seek personal wealth or generate it for others, but instead, to contribute in some way to a greater cause or develop a tool that can benefit those who truly need it.

For me, true satisfaction lies in being able to use my skills to make a positive difference in the lives of others. I would love to be able to add value and strive to improve the quality of life for a particular group through my work in Data Analysis. My intention is to use my knowledge to the best of my ability, hoping to make a small yet significant contribution to society.

Leaving personal motivations aside, the significance of undertaking this proect becomes apparent given the current situation of chronic diseases. As can be observed in the following graph based on data extracted from the Instituto Nacional de Estadística (INE, 2022), the majority of the leading causes of mortality in Spain are classified as chronic diseases.

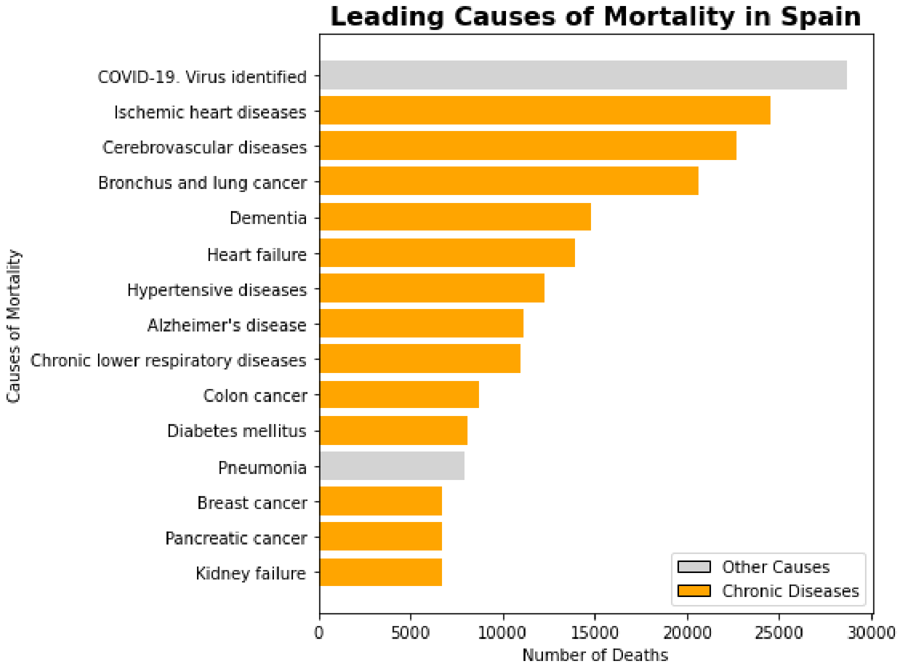


Figure 1.1: Leading Causes of Mortality in Spain

While reducing mortality from chronic diseases has been the primary focus of global strategies and goals in this field, these diseases are also the main source of disability and multimorbidity. Currently, 80% of years lived with a disability worldwide are caused by chronic diseases (Alliance, 2022). These disabilities have a significant impact on individuals' Quality of Life, limiting their autonomy and participation in society.

On top of everything mentioned so far, it is worth considering whether governments are aware of this situation, as 1 in 5 people live with at least one chronic disease, yet only 0.6%-1.6% of resources allocated to supporting the development and strengthening of healthcare systems are dedicated to chronic diseases (Alliance, 2022).

The Quality of Life of individuals affected by this situation is typically assessed either through in-person interviews with a healthcare professional or through self-administered questionnaires, both occasionally supplemented with patients' testimonials about their experiences with the disease.

This methodology is highly time and resource-consuming, both for healthcare professionals and the patients themselves. It is in this context that the automation of certain processes and the use of advanced technologies could play a crucial role. By implementing systems and tools that allow for a more efficient and precise assessment of quality of life, healthcare appointments would be relieved and freed up, enabling professionals to dedicate more time to direct patient care and clinical decision-making.

1.2 Objectives

Taking into account the facts presented in the previous section, the importance and urgency of addressing the topic of chronic diseases and Quality of Life from a Data Analysis perspective becomes evident. In this regard, the present study aims to achieve the following objectives.

The main objectives of this study are:

- Development a model that identifies features or indicators of the Quality of Life of patients with chronic diseases given a textual testimony, achieving an acceptable performance.
- Conduct a comparative analysis of different architectures and techniques for constructing such model.
- Undertake a significant initial approach to the proposed problem and lay the groundwork for a larger research project to be taken on by the ProHealth group.

In terms of the student's academic development, there are other objectives that could be considered secondary:

- Improve the student's formation as a Data Scientist through participation in a real project alongside a multidisciplinary team composed of professionals from various fields.

- Provide a significant introduction to the field of Natural Language Processing, which, despite being one of the prominent areas in the current landscape of Artificial Intelligence, is not covered in any subject within the Master's program.
- Strengthen knowledge originally obtained from numerous subjects in the Master's program such as Data Mining, Data Visualization, Machine Learning, and Cloud Computing, which will play an important role in the project's development.

1.3 Document Structure

This document is composed of 5 chapters, two of which, "Theoretical Framework" and "Methodology," cover a significant portion of its content. The following lines provide an overview of what will be covered in each chapter, allowing the reader to anticipate the document's content in broad terms.

As can be observed, **Chapter 1: Introduction**, is dedicated to introducing the study and justifying its relevance. The objectives to be achieved through its elaboration are outlined, and the structure that will be followed to accomplish them is explained.

Chapter 2: Theoretical Framework is a crucial chapter as it addresses the study's underlying theoretical aspects. First, the field of Quality of Life and chronic diseases is introduced, including multiple indicators, methodologies, and tools for their evaluation. Next, we provide an overview of Natural Language Processing (NLP), situating the concept within the broader context of Artificial Intelligence. We discuss the main tasks within NLP and explore the different phases of an NLP-based project. Finally, we review the state of the art regarding the application of NLP techniques in the evaluation of Quality of Life for individuals with chronic diseases.

In **Chapter 3: Methodology**, also quite extensive, we describe the available labelled data and how we process it, as well as our approach to developing Machine Learning models to achieve the objectives set in Chapter 1. Additionally, we explain the various tools required to carry out different parts of this development.

The results obtained from the process outlined in Chapter 3 are presented in **Chapter 4: Results**. Here, we compare the performance achieved through different approaches to the proposed problem, based on predefined metrics.

Finally, **Chapter 5: Conclusions** highlights the main findings from this study. It also addresses the limitations encountered and suggests multiple lines for future work, considering the potential continuation of this project.

2. Theoretical Framework

In this chapter, we will address the underlying theoretical framework that supports the present study. We will provide the conceptual and theoretical foundations upon which the research is based, offering a comprehensive overview of relevant theories, models, and studies in the research area.

2.1 Chronic Diseases and Quality of Life

As per the World Health Organization (WHO), noncommunicable diseases (NCDs), also known as chronic diseases, tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behavioural factors. The main types of NCD are cardiovascular diseases (such as heart attacks and stroke), cancers, chronic respiratory diseases (such as chronic obstructive pulmonary disease and asthma) and diabetes. NCDs disproportionately affect people in low- and middle-income countries, where more than three quarters of global NCD deaths (31.4 million) occur (WHO, 2022b).

In the present, there is an increasing fraction of the population affected by chronic diseases. These conditions, characterized by their long-lasting nature and slow progression, have become the leading cause of mortality and disability worldwide. Consequently, they pose significant social and health challenges on a global scale. Detection, screening and treatment of NCDs, as well as palliative care, are key components of the response to theese diseases (WHO, 2022a).

The majority of chronic diseases hold the potential to worsen the overall health of patients by limiting their capacity to live well, limit the functional status, productivity and Quality of Life, and are a major contributor to health care costs. A chronic disease disrupts an individual's life and this disruption may be interpreted in terms of its impact on well-being or Quality of Life (Devins et al., 1983).

Different types of diseases can result in limitations in one or multiple aspects that impact Quality of Life. These can include physical functioning, mobility, cognitive abilities, social relationships, emotional well-being, satisfaction with life, work, education, and economic stability.

Now, let us delve into the concept of Quality of Life. According to WHO, Quality of Life (QoL) is defined as an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns (Bonomi et al., 2000). QoL is the feeling of overall life satisfaction, as determined by the mentally alert individual whose life is being evaluated (Meeberg, 1993) .

There is a general consensus that Health Related Quality of Life encompasses multiple dimensions, including physical, psychological, and social functioning, which are impacted by an individual's disease and/or treatment. Physical functioning pertains to the ability to carry out various activities of daily living and encompasses both the physical symptoms arising from the disease or treatment. Psychological functioning encompasses a range from significant psychological distress to a positive sense of well-being, and may also include cognitive abilities. Social functioning relates to both the quantitative and qualitative aspects of social relationships, interactions, and integration into society (Sprangers, 2002).

Depending on the individual and his particular situation, these different aspects of life affected by an illness will be reacted to and assumed in a wide variety of manners (Lacroix et al., 1995). Understanding the impact of chronic illness on an individual's overall well-being and daily functioning is crucial for healthcare professionals in order to provide comprehensive care and support tailored to the unique needs of each patient.

Health related QoL can be assessed either by interview or questionnaire. Interview methods use semi-structured approaches, which are useful for the initial creation of items to be subsequently used in questionnaires to discover issues and describe the experiences of the patients (Aaronson, 1989). This assesment can also be completed with subjective information and experiences captured by patient testimonials.

Patient testimonial refers to the written or spoken narrative provided by a patient about their experiences and perspective regarding their illness or medical condition. Theese play a significant role since, unlike interviews and questionnaires, they offer a rich, personal and unstructured narrative that delves into the lived experiences, emotions, and perspectives of individuals facing chronic illnesses. Testimonials offer unique insights into the day-to-day challenges, coping mechanisms, and personal journeys of patients, allowing researchers and healthcare professionals to gain a deeper understanding of the impact of illness on various aspects of life.

These free-text comments are especially useful if they are reported and analized with the same scientific rigor as the closed-ended questions. Unstructured free-text patient comments contain valuable information, but the manual analysis of this kind of data requires a large number of resources and personnel that are not available in most healthcare organizations. Thus, this process is limited by the absence of a system to systematically extract information of the free-text patient comments that will improve the quality of care (Barber et al., 2021; Cognetta-Rieke & Guney, 2014; Forsyth et al., 2018; Khanbhai et al., 2021).

The ultimate objective of this process is to measure patients's Quality of Life since it is important to evaluate the impact of their chronic disease, and physiological measurements are often poorly correlated with functional capacity and well-being. Another reason to measure QoL is the commonly observed phenomena that two patients with the same clinical criteria often have dramatically different responses. For example, two patients with the same range of motion and even similar ratings of back pain may have different role function and emotional wellbeing. Although some patients may continue to work without major depression, others may quit their jobs and have major depression (Guyatt et al., 1993).

National accounts aggregates, such as Gross Domestic Product (GDP), have become crucial indicators for assessing economic performance and living standards. These aggregates enable easy comparisons and serve as a widely used measure of economic activity. Decision-makers and policy-makers often rely on GDP as a benchmark for their actions and recommendations. While GDP encompasses all final goods and services produced by an economy, it was not originally designed as a measure of social progress, although historically it has been considered to be closely linked to the well-being of citizens (Eurostat, 2017).

However, GDP is an inadequate measure of citizens' quality of life due to its limitations: it fails to reflect households' true economic situations, disregards the depletion of resources for future generations, and overlooks wealth distribution among the population. Relying solely on GDP as a metric can lead to an incomplete understanding of citizens' well-being (Eurostat, 2017).

Within the alternative forms of Quality of Life assesment needed to complement GDP in covering other important domains related to well-being, we encounter what are commonly referred to as Quality of Life Indicators. An indicator is characterized as an observable variable assumed to point to, or estimate, some other (usually unobservable) variable (Bunge, 1975).

It is usual to consider that indicators may be objective or subjective. This distinction is very relevant in the sense that both objective situations and subjective perceptions determine in fine the well-being of the person. The consensus that both dimensions matter and should be measured, is now accepted. Some indicators are clearly objective, like the income. Others are clearly subjective, like life satisfaction. However, the border between subjective and objective indicators is often blurred by the measurement methods (Eurostat, 2017).

Quality of Life Assessment Instruments (Questionnaires)

Self- or interviewer-administered questionnaires can be used to measure cross-sectional differences in quality of life between patients at a point in time (discriminative instruments) or longitudinal changes in HRQoL within patients during a period of time (evaluative instruments). Both discriminative and evaluative instruments must possess validity, ensuring an accurate measure of the targeted aspects, as well as high levels of reliability and responsiveness (Guyatt et al., 1993). Health-related Quality of Life measures should also be interpretable alowing clinicians to identify differences in scores.

Quality of Life measurement approaches are divided in two basic groups: generic instruments that provide a summary of health related QoL or specific instruments that focus on problems associated with a single disease. Each approach has its strengths and weaknesses and may be suitable for different circumstances. However, both these instruments are suitable for detecting minimally important effects in clinical trials (Guyatt et al., 1993).

Some of the most commonly used scales and questionnaires to assess HRQoL include the WHOQOL (World Health Organization Quality of Life) (The Whoqol Group, 1998), the EuroQol-5D (EuroQol Group, 1990), and the SF-36 (Short Form-36) (Ware & Sherbourne, 1992). These scales vary in terms of the number and nature of the items assessed, as well as in how the results are scored and analyzed. The latter, SF-36, will be further examined in detail, as later on we will see it playing a pivotal role in this study.

- **WHQOOL:** Developed by the World Health Organization, it is a sophisticated assessment tool that evaluates individuals' subjective well-being across multiple dimensions, including physical health, psychological well-being, social relationships, and environmental factors. It provides valuable insights into the holistic nature of Quality of Life, informing health interventions to enhance overall well-being (The Whoqol Group, 1998).

- **EuroQoL-5D:** It is a widely used instrument for assessing health-related Quality of Life. It was developed by the EuroQol Group, an international collaborative effort. The questionnaire measures Quality of Life across five dimensions (5D): mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has multiple levels or response options, allowing individuals to indicate their level of functioning or satisfaction in each area. By capturing these dimensions, the EQ-5D provides a comprehensive perspective on health-related Quality of Life, enabling researchers and clinicians to compare individuals' overall well-being (EuroQol Group, 1990).

- **SF-36:** It is a comprehensive measure that captures multiple dimensions of well-being, including physical functioning, role limitations, bodily pain, general health perceptions, vitality, social functioning, emotional well-being, and mental health. This instrument, developed by (Ware & Sherbourne, 1992), provides valuable insights into various aspects of an individual's health status. Since it will serve as a foundation for further exploration in our study, we shall now take a deep dive and revise it more carefully.

SF-36

> *"The SF-36 is a 36-item questionnaire that measures health-related quality of life across eight domains: physical functioning, role limitations due to physical health, bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems, and mental health. It was developed as part of the Medical Outcomes Study (MOS) and has been widely used in research and clinical practice to assess patient outcomes and inform healthcare decision-making"* (Ware & Sherbourne, 1992).

The eight domains measured by the SF-36 are:

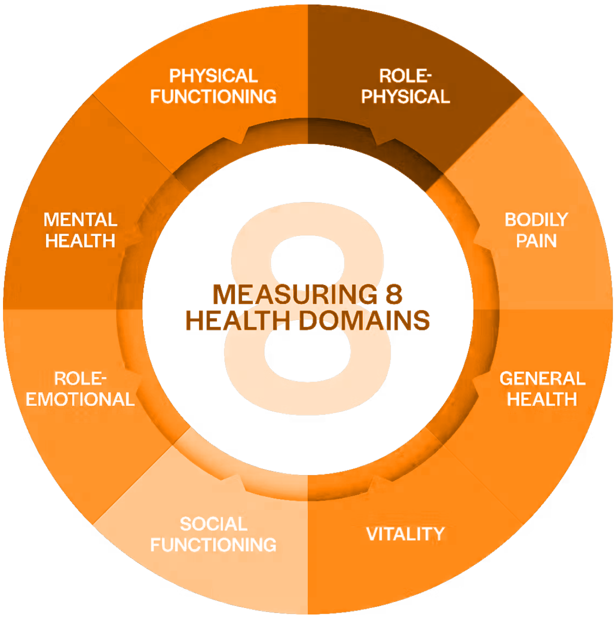


Figure 2.1: SF-36 domains (Fleishman et al., 2006)

- **Physical functioning:** This domain assesses the ability to perform physical activities such as walking, climbing stairs, and carrying groceries.
- **Role limitations due to physical health:** This domain measures the extent to which physical health problems interfere with work or other daily activities.
- **Bodily pain:** This domain assesses the severity and impact of pain on daily activities.

- **General health perceptions:** This domain measures overall perceptions of health, including energy level, health outlook, and resistance to illness.
- **Vitality:** This domain assesses energy level and fatigue.
- **Social functioning:** This domain measures the ability to perform social activities and the extent to which physical or emotional problems interfere with social activities.
- **Role limitations due to emotional problems:** This domain measures the extent to which emotional problems interfere with work or other daily activities.
- **Mental health:** This domain assesses psychological distress, well-being, and overall mental health.

Each domain is measured by a subset of questions, with a total of 36 questions across all domains. The questions are answered on a Likert scale, with response options ranging from "all of the time" to "none of the time" or "extremely" to "not at all." The responses are then scored and transformed to a scale from 0 to 100, with higher scores indicating better health-related quality of life.

(A Likert scale consists of a series of statements or items to which respondents indicate their level of agreement or disagreement on a predetermined scale.)

The result of a SF-36 questionnaire might be represented as follows:

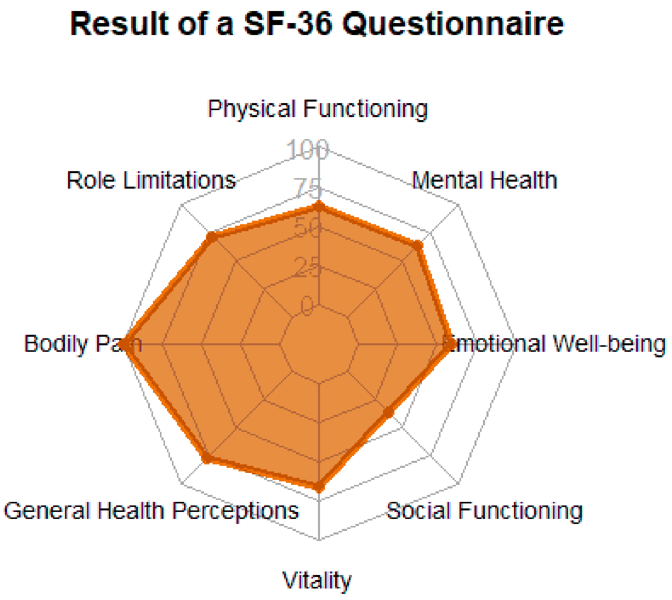


Figure 2.2: Simulated SF-36 result

2.2 Natural Language Processing

Natural Language Processing (NLP) is a field at the intersection of Computer Science, Artificial Intelligence, and Linguistics. It concerns building systems that can process and understand human language. Since its inception in the 1950s and until very recently, NLP has primarily been the domain of academia and research labs, requiring long formal education and training. The past decades breakthroughs have resulted in NLP being increasingly used in a range of diverse domains such as retail, healthcare, finance, law, marketing, human resources, and many more.

In the following lines, a brief introduction to Natural Language Processing will be provided, covering its basic concepts, some of its most advanced techniques, and the current state of the art.

By natural language we mean a language that is used for everyday communication by humans; languages such as English, Hindi, or Spanish. In contrast to artificial languages such as programming languages and mathematical notations, natural languages have evolved as they pass from generation to generation, and are hard to pin down with explicit rules. We will take Natural Language Processing in a wide sense to cover any kind of computer manipulation of natural language (Bird et al., 2009).

To understand how NLP fits into the broader landscape of Artificial Intelligence, let us revisit the definitions of the following disciplines:

Loosely speaking, Artificial Intelligence (AI) is a branch of Computer Science that aims to build systems that can perform tasks that require human intelligence. The foundations of AI were laid in the 1950s and initial AI was largely built out of logic, heuristics, and rule-based systems (Vajjala et al., 2020). However, AI has evolved significantly since then, driven by advancements in computing power and data availability. It has transformed numerous industries and achieved human-level performance in sectors like healthcare, finance, and transportation, shaping our daily lives.

Machine Learning (ML) is a branch of AI that deals with the development of algorithms that can learn to perform tasks automatically based on a large number of examples, without requiring handcrafted rules. From the efforts of mega corporations such as Google, Microsoft, Facebook, Amazon, and so on, Machine Learning has become one of the hottest computational science topics in the last decade. (Edgar & Manz, 2017; Vajjala et al., 2020).

Deep Learning (DL), a subset of ML, is inspired by the information processing patterns found in the human brain. It uses a large amount of data to map the given input to specific labels and is designed using numerous layers of algorithms (artificial neural networks), each of which provides a different interpretation of the data that has been fed to them (LeCun et al., 2015; Vajjala et al., 2020).

ML, DL, and NLP are all subfields within AI, and the relationship between them is depicted in the following figure:

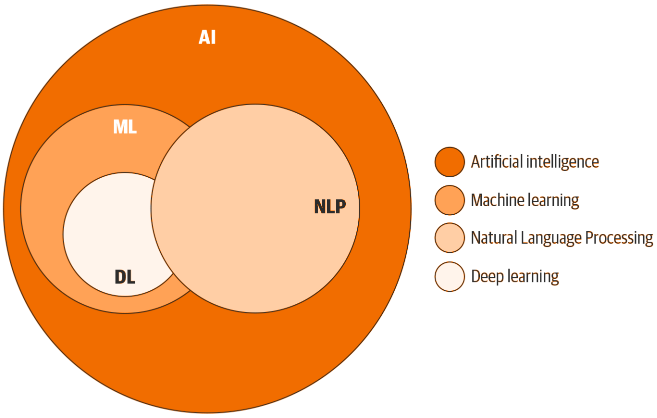


Figure 2.3: AI, ML, DL and NLP relationship (Vajjala et al., 2020)

Like other early work in AI, early NLP applications were also based on rules and heuristics. In the past few decades, though, NLP application development has been heavily influenced by Machine Learning based methods.

Once the general landscape of NLP and its relationship with other subdisciplines of Artificial Intelligence has been understood, we will dive into the technical aspects and formally explore how Natural Language Processing works and its main applications.

According to estimates by IDC (International Data Corporation), unstructured data was projected to account for 95% of global data in 2020, with an annual growth rate of approximately 65% (Adnan & Akbar, 2019). As a significant portion of this data exists in the form of free-text, the utilization of NLP becomes crucial. NLP models are capable of performing a wide range of tasks with varying levels of difficulty. The following are among the most common of them:

- **Infromation Retrieval:** Information retrieval (IR) systems, such as search engines, map an information need to a written document or list of documents sorted by a measure specific to the goals of the system. This measure is typically a combination of relevance to a need, information quality, and business goals (McRoy & Ali, 2021). The goal is to retrieve the most relevant documents that match the user's information needs.
- **Information Extraction:** Information extraction (IE) involves identifying entities within unstructured text and organizing them into (predefined) structures that can be used as input to downstream tasks (McRoy & Ali, 2021).

- **Document Classification:** Document classification is the task of assigning predefined cathegories to a document based on its content.
- **Question Answering:** Question Answering (QA), also known as information-seeking dialog, is a type of two-way interaction where the user controls the primary turns and the system responds. Such a system must first interpret the question and then create and deliver the response (McRoy & Ali, 2021).
- **Language Generation:** Natural Language Generation (NGL) is defined as the subfield of Computational Linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying linguistic or non-linguistic representation of information (Gatt & Krahmer, 2018).

In this study, since the focus will be on extracting quality of life indicators present in chronically ill patients' testimonies, we will be in the domain of Information Extraction. The task involves identifying and extracting specific indicators related to quality of life from unstructured text. Information Retrieval, on the other hand, would focus more on retrieving relevant testimonies based on specific queries or criteria, rather than extracting specific indicators from the text.

Information Extraction Systems in Clinical applications can be cathegorized in two different groups: General Purpose IE Systems, which aim to extract all types of information (diseases, drugs, lab tests, procedure ...) and Specific Purpose IE Systems, which, conversely, are more focused on certain information and seek to have a better performance in specific tasks (Keloth et al., 2023).

Going into further detail about Information Extraction, it is commonly understood that it consists of three main components (Keloth et al., 2023):

- **Named Entity Recognition (NER):** Involves identifying and classifying specific entities or elements in a text. These entities can include person names, organization names, locations, dates, and other similar information. The objective of NER is to automatically identify and label these named entities providing valuable insights and understanding.
- **Relation Extraction:** Extraction of modifiers to the main entities, such as negation, subject, conditional, certainty or temporality.
- **Concept Normalization:** Also known as Entity Linking, is the action of linking entities to concepts in an ontology (formal representation of knowledge within a specific domain).

In general, NLP models work by finding relationships between the constituent parts of language, for example, the letters, words, and sentences found in a text dataset (Deeplearning.ai, 2023). This process is usually divided into 3 main parts: Data Processing, Feature Extraction and Modelling.

Data Preprocessing

As it is in every AI application, data preprocessing is a crucial step in NLP, since the quality of the data can highly influence the overall performance of the models.There are a series of NLP-specific transformations that are widely used and that we will cover next:

NLP software typically analyzes text by breaking it up into words (tokens) and sentences. Hence, any NLP pipeline has to start with a reliable system to split the text (Deeplearning.ai, 2023; Vajjala et al., 2020).

- **Sentence Segmentation:** Breaks a large piece of text into linguistically meaningful sentence units. This is obvious in languages like English, where the end of a sentence is marked by a period, but it is still not trivial since the period symbol also has other uses.
- **Word Tokenization:** Similar to the previous one, in this case, the text is divided into words based on ”space“ characters.
- **Punctuation Removal:** Removing punctuation and/or numbers is also a common step for many NLP problems.
- **Lowercase:** Upper or lowercase normally do not make a difference for the problem. Therfore, all text is lowercased (or uppercased, although lowercasing is more common).
- **StopWord Removal:** Some of the frequently used words in English (and other languages), such as ”a“, ”an“, ”the“, ”of“, ”in“, etc., are not particularly useful for NLP tasks, as they dont carry any content on their own. Such words are called stop words and are typically (though not always) removed from further analysis. There is no standard list of stop words for English, although there are some popular lists. Bear in mind that what a stop word is can vary depending on the project at hand.
- **Stemming and Lemmatization:** Stemming is an informal process of converting words to their base forms using heuristic rules, specifically it’s the process of removing suffixes and reducing a word to some base form such that all different variants of that word can be represented by the same form. One limitation in this approach is that different words may be mapped to the same base form, even though they dont have a close semantic relationship.

  On the other hand, Lemmatization is the process of mapping all the different forms of a word to its lemma. Lemmatization requires more linguistic knowledge, and modeling and developing efficient lemmatizers remains an open problem in NLP research even now.

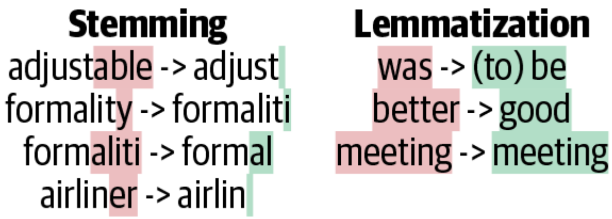


Figure 2.4: Difference between Stemming and Lematization (Vajjala et al., 2020)

There are also some other advanced data preprocessing techniques that fall outside the scope of the current work, such as POS Tagging, Parsing and Coreference Resolution.

Feature Extraction and Modelling

The Machine Learning models we intend to apply in the Modelling stage will require us to convert all this textual information into numerical values. The process by which this conversion of sentences and words into numerical values is performed is known as Feature Extraction, Feature Engineering, or Text Representation.

In the field of NLP, there exists a close interrelation between the stages of Feature Extraction and Modelling. In many cases, the process of extracting features is carried out as part of the training of the selected model, making it difficult to establish a clear separation between these stages.

Instead of viewing them as distinct and sequential stages, it is more appropriate to perceive them as interdependent processes that merge in the construction of the NLP model. During training, the NLP model learns to automatically extract relevant features from the text, including the representation of words and phrases, as well as capturing linguistic relations.

Two main approaches are commonly used for text representation in NLP, the Classical Machine Learning approach or the Deep Learning-based approach.

All the feature extraction techniques we will mention have the main objective of transforming words or texts into numerical vectors in a vector space. This transformation is necessary for Machine Learning models to process and work with text data, as they require numerical representations. In the following lines, we will explore various approaches to feature extraction, and through different examples, we will gain intuition behind the representation of text in vector spaces. Next, we will mention some interesting mathematical properties that we may encounter.

Classical ML Approach:

These techniques rely on word counting or term frequency in a document to create numerical features that represent the text.

- **Bag-of-Words:** A Bag-of-Words (BoW) is a representation of text that describes the occurrence of words within a document. It involves a vocabulary of known words and a measure of the presence of them. In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature, all the structure and linear ordering of words within the context is ignored (Goldberg, 2017; Manning & Schütze, 1999).

  In the case where a word appears four times in a document, the corresponding position in the feature vector will have a count of 4. Conversely, if a word from the vocabulary does not appear in the document, it will be assigned a count of 0. Let us consider a simple example to illustrate the interpretation of the vector representation.

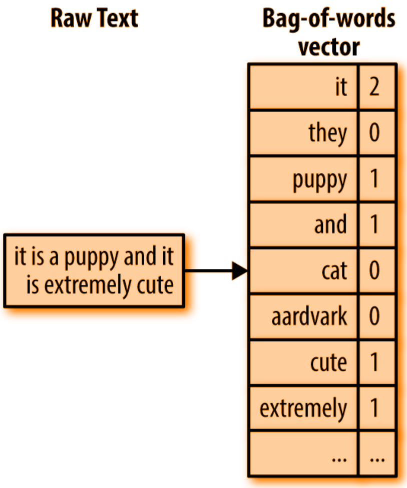


Figure 2.5: BoW Example (1/3) (Soo, 2023)

  The original text is a sequence but the BoW has no sequence, it just captures how many times each word appears in the text. In a BoW vector, each word becomes a dimension of the vector. If there are n words in the vocabulary, then a document becomes a point in a n-dimensional space (Soo, 2023). To grasp this representation, some 2 and 3 dimensional simple examples are shown in the following images.

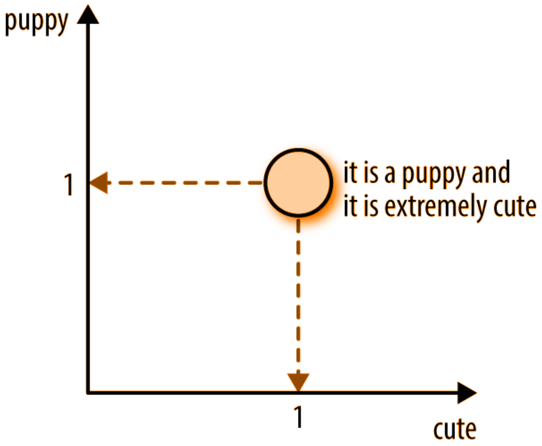


Figure 2.6: BoW Example (2/3) (Soo, 2023)

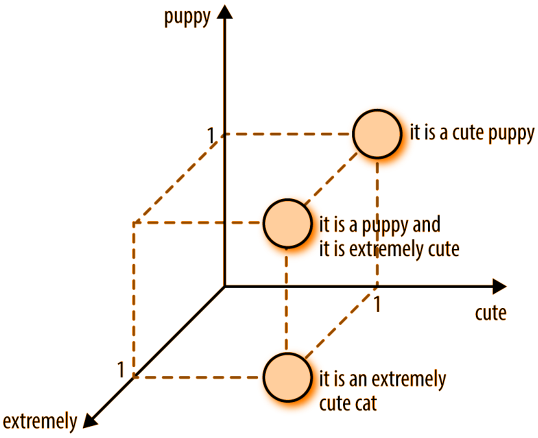


Figure 2.7: BoW Example (3/3) (Soo, 2023)

The first image shows what our example sentence looks like in the two-dimensional feature space corresponding to the words puppy and cute. The second image shows three sentences in a 3D space corresponding to the words puppy, extremely, and cute (Soo, 2023).

- **TF-IDF:** The Bag-of-Words representation is straightforward to generate, however, it has limitations. By assigning equal weight to all words, some receive more emphasis than necessary. Ideally, we seek a representation that accentuates meaningful words in a text.

  TF-IDF is a slight modification of the bag-of-words approach, known as term frequency inverse document frequency. Instead of considering the raw counts of each word in every document of a dataset, TF-IDF calculates a normalized count by dividing the word count by the number of documents in which the word appears. This can be formally expressed as:

$$\textit{bow}(w,d) = \textit{ count of word } w \textit{ in document } d \tag{2.1}$$

$$\textit{tf-idf}(w,d) = \textit{bow}(w,d) \cdot \frac{N}{\textit{count of docs in which } w \textit{ appears}} \tag{2.2}$$

  The fraction shown in (2.2) is known as inverse document frequency. If a word is present in numerous documents, its inverse document frequency approaches 1. Oppositely, if a word appears in only a reduced number of documents, its inverse document frequency is significantly higher.

In the BoW model, the vector size corresponds to the vocabulary size. When the majority of vector values are zero, the BoW representation becomes a sparse matrix, and theese pose challenges for both computational efficiency and information handling.

While the TF-IDF model contains the information on the more important words and the less important ones, it does not solve the challenge of high dimensionality and sparsity, and, similar to BoW, it also makes no use of semantic similarities between words.

DL-based Approach:

These Deep Learning-based techniques solve the previously mentioned problems by introducing what are known as Word Embeddings.

A word embedding is a function that maps each word type to a single vector. These vectors are typically dense and have much lower dimensionality than the size of the vocabulary. In general, words that have similar meaning will have very similar vector representations.

These word embeddings are generated using techniques like Word2Vec or GloVe (Mikolov, Chen, et al., 2013; Pennington et al., 2014), which learn word representations based on the surrounding words or co-occurrence patterns. We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language (Mikolov, Yih, et al., 2013).

The word representations computed using neural networks are very interesting because the learned vectors explicitly encode many linguistic regularities and patterns. Somewhat surprisingly, vector spaces created by word embedding representations have certain mathematical properties that enable many of these patterns to be represented as linear translations.

For example, the the famous example introduced by (Mikolov, Sutskever, et al., 2013): the vector calculation

$$vec(\text{``}Madrid\text{''}) - vec(\text{``}Spain\text{''}) + vec(\text{``}France\text{''})$$

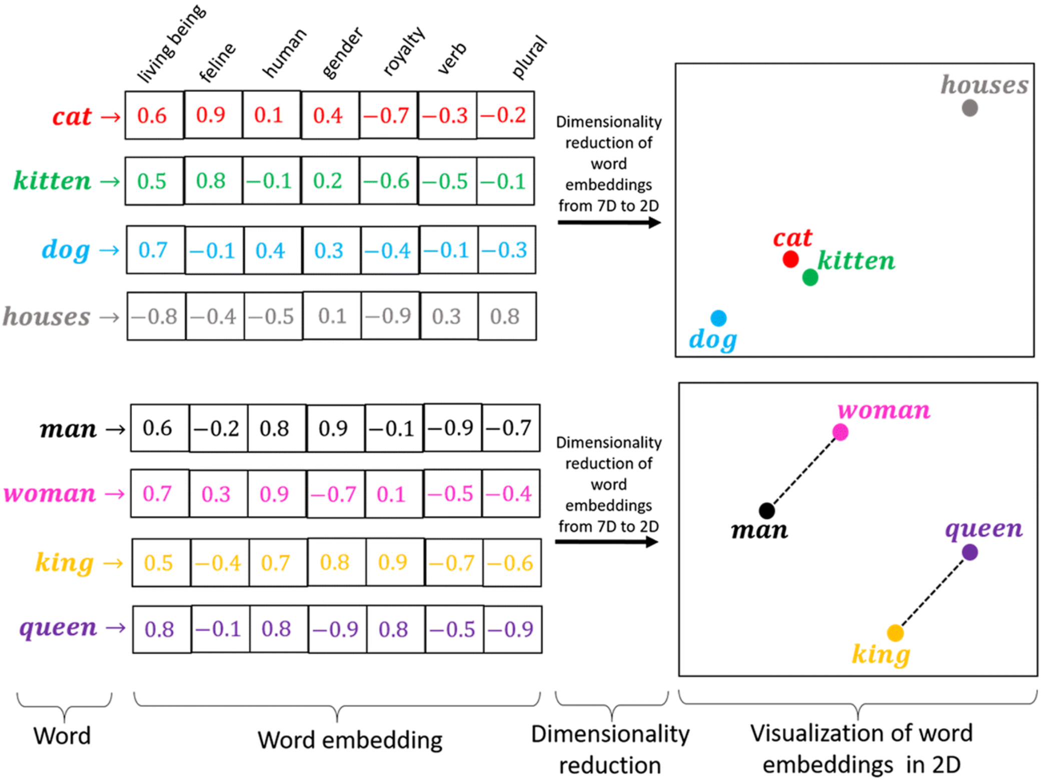
is closer to $vec(\text{``}Paris\text{''})$ than to any other word.

Another famous example that was also introduced by (Mikolov, Yih, et al., 2013) is that by subtracting the vector representation of "man" from "king," we obtain a vector that captures the semantic difference between the two words. Then, by adding the vector representation of "woman" to the result, we find a vector that is close to the vector representation of "queen":

$$vec(\text{''}King\text{''}) - vec(\text{''}Man\text{''}) + vec(\text{''}Woman\text{''}) \approx vec(\text{''}Queen\text{''})$$

As it can be observed in (Le & Mikolov, 2014), the concept of word embedding is not only applicable to individual words, but has also been successfully extended to phrase and document-level. This expansion to larger text units has enabled capturing more intricate semantic relationships, opening up new possibilities in the field of text comprehension and generation.

Below, we present an illustrative instance of a word embedding, showcasing example vectors and the corresponding word representations in a vector space reduced to two dimensions:



	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Figure 2.8: Word Embedding Example (Soo, 2023)

Now that we have familiarized ourselves with the concept of word embeddings, let us briefly introduce the most significant ones:

- **Word2Vec:** The Word2Vec model is based on the combination of CBOW (Continuous Bag of Words) and Skipgram. The former utilizes a shallow neural network and a fixed-size sliding window to predict each word in a document given its context. Through iterative training on a large corpus of text, the neural network weights are progressively refined. The Skipgram approach is very similar, but inverted: given the central word in a sliding window, a neural network is trained to predict the surrounding words (context). Again, this iterative process refines the neural network weights. The combination of these weights, multiplied by a OneHotEncoded vector for each word in the vocabulary, allows us to obtain the embedding of each word, that is, its numerical vector representation in a high-dimensionality vector space (Mikolov, Chen, et al., 2013).
- **GloVe:** The GloVe (Global Vectors) model is a sophisticated approach for learning word representations that leverages the statistics of word occurrences

in a corpus to capture the meaning of words. By analyzing word-word co-occurrence counts, the model constructs a matrix that represents the relationships between words. This matrix is then used to calculate the probability of a word appearing in the context of another word.

The key innovation of the GloVe model is its ability to capture both the frequency-based information and the linear patterns observed in previous prediction-based methods like Word2Vec. This allows GloVe to learn word representations that excel in tasks such as word analogy, word similarity, and named entity recognition (Pennington et al., 2014).

- **Transformer-based Models:** Before introducing a few Transformer-based models, it is necessary to provide a brief overview of what Transformers are. In the highly influential article "Attention is All You Need" Ashish Vaswani et al. from Google introduce the Transformer, a groundbreaking neural network architecture that revolutionizes the field of sequence modeling.

  > *The key idea behind the Transformer is to enable the model to capture global dependencies between input and output sequences. This is achieved through a mechanism called self-attention, where each position in the input sequence can attend to all other positions to compute a weighted representation. This allows the model to consider the context and relevance of each input position when generating the output.*
  >
  > *One of the major advantages of the Transformer is its ability to parallelize computations. Traditional recurrent models process sequences sequentially, limiting parallelization. In contrast, the Transformer can process all input positions simultaneously, resulting in significantly faster training times. Additionally, the Transformer addresses the challenge of learning dependencies between distant positions in a sequence, which is a problem that makes traditional recurrent networks struggle* (Vaswani et al., 2017).

  With that being said, let's now introduce two of the most well-known transformer-based models that are leading the current "revolution" in the field of Artificial Intelligence. These models have amazed the world, bringing the extraordinary power of modern AI to everyone's fingertips.

  - **BERT:** Stands for Bidirectional Encoder Representations from Transformers, and is a language representation model developed by Google that is pre-trained on large amounts of unlabeled text data, allowing it to learn rich, contextualized representations of words and sentences. Unlike traditional language models that process text in a unidirectional manner, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. This in simple terms means that BERT, unlike humans, processes input both left to right and right to left.

BERT is based on the Transformer architecture where the encoder is used to pre-train the model on large amounts of unlabeled text data, allowing it to learn rich, contextualized representations of words and sentences. These pre-trained representations can then be fine-tuned for specific NLP tasks by adding a task-specific output layer to the model and training it on labeled data (Devlin et al., 2019).

- **GPT:** At its core, GPT (Generative Pre-training) leverages the power of unsupervised learning on vast amounts of unlabeled text data to pre-train a language model. During pre-training, GPT employs a generative modeling approach, wherein the model learns to predict the next word in a sentence given the preceding context. By training on a multitude of sentences, the model gradually acquires a deep understanding of the syntactic, semantic, and contextual aspects of language.

  Once the pre-training phase is complete, GPT undergoes a fine-tuning process to adapt its learned representations to specific downstream tasks. This supervised fine-tuning involves training the model on labeled data for the target task, such as question answering or document classification. What sets GPT apart from previous approaches is its ability to effectively transfer the knowledge gained during pre-training to the target task (Radford, Narasimhan, et al., 2018; Radford, Wu, et al., 2018).

  OpenAI has made progressive improvements to the GPT models, culminating in the most recent iteration known as GPT-4. This latest version marks a significant advancement as it introduces multimodal capacity, enabling the model to process diverse data formats including images. Furthermore, GPT-4 showcases improved memory, language comprehension, and contextual understanding, thanks to its remarkable inclusion of over a trillion parameters.

Both BERT and GPT were introduced in 2018 by Google and OpenAI, respectively. Although both models are unsupervised learning models that use the Transformer architecture to learn context from textual-based datasets using attention mechanisms, they differ in multiple aspects.

While BERT is bidirectional, meaning it takes into account both left and right context, GPT-3 is an autoregressive model, only considering the left context when making predictions. This makes BERT particularly well-suited for tasks such as sentiment analysis and natural language understanding.

Moreover, it is worth mentioning that GPT and BERT were trained on different sizes of text datasets. GPT utilized a vast dataset of 45TB, while BERT was trained on a relatively smaller dataset of 3TB. This difference in training data volume might give GPT an advantage in tasks like translation or summarization.

Developing models like GPT or BERT, which require training millions of parameters, entails an extremely high computational cost. It is not feasible to expect that every user or organization intending to utilize word embeddings has to go through a such resource-intensive training.

For this reason, using pre-trained embeddings has become a common practice for users and organizations lacking sufficient computing capabilities. By leveraging these pre-existing embeddings, which are trained on extensive datasets, they can benefit from the expertise and resources of larger entities.

This approach offers several advantages, including efficient resource utilization, generalization to various tasks, transfer learning capabilities, and time-saving benefits. Users can fine-tune the pre-trained embeddings to their specific domain or task, achieving better performance with limited labeled data.

With the aforementioned considerations in mind, note that embeedings are not exclusive to the NLP domain since in recent times some innovative tools like Meta's ImageBind (Girdhar et al., 2023) have emerged. These tools rely on multimodality (training models with different types of data) and aim to generate joint embeddings where different representations of the same concept are mapped to nearby points of a shared vector space. For instance, the texts "a bird flying" and "a bird gliding" should have close representations, as well as an image of a seagull in flight and the sound of bird wings flapping.

This allows the extension of the previously mentioned word embedding arithmetic to multiple data types allowing some "calculations" as the following:

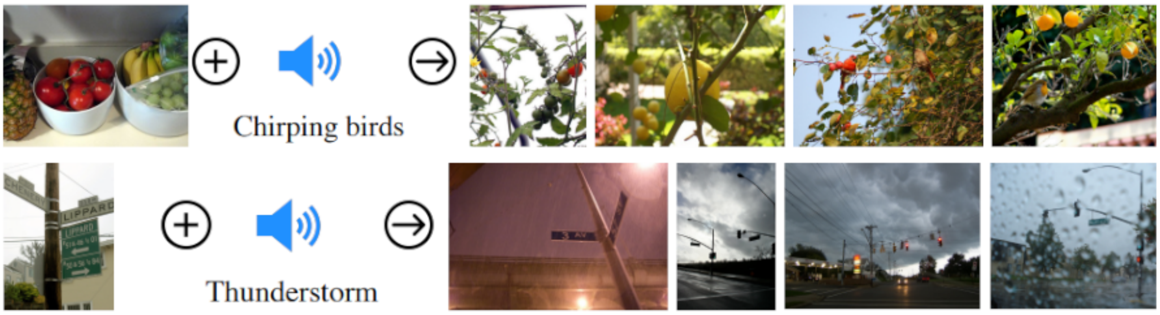


Figure 2.9: Joint Embedding Arithmetic Examples (Girdhar et al., 2023)

2.3 NLP in QoL assessment

In this study, we aim to address the combination of the two aforementioned sections, contributing to the application of NLP in assessing Quality of Life. In the present section, we will examine the current landscape and state of the art in this field.

The application of NLP in the field of healthcare is a promising idea due to the vast amount of data available in healthcare organizations. This data, in the form of free-text, contains valuable information related to patient care and outcomes.

As mentioned earlier, the manual analysis of this unstructured clinical text data demands significant resources and personnel, which are often not accessible in the majority of healthcare organizations. A solution to mitigate theese limitations is the use of Artificial Intelligence. AI has a relevant application in the health area and plays an increasingly important role in the area of biomedicine (Escrivá et al., 2020). In this context, machine learning and, specifically, natural language processing (NLP) are going to be the critical methodology for processing unstructured free text.

Nowadays, Big Data and data mining technologies are increasingly present in the healthcare field. They make it possible to analize, transform, synthesize and provide intelligence to the data, and transform the large amount of data into knowledge that supports diagnosis and decision making. NLP is used to extract information, convert unstructured text into a structured format, perform syntactic processing, capture meaning and identify relationships between concepts (Klein et al., 2020; Lindvall et al., 2018; Robert & Cornwell, 2013; Velupillai et al., 2018). NLP can be used in various scientific fields and for myriad purposes such as linguistic analysis, information retrieval, text translation, conversational bots, text classification, and human emotion analysis, among others.

One notable project in this domain is the Linguistic String Project-Medical Language Processor (LSP-MLP). The LSP-MLP aids physicians in extracting and summarizing information related to signs, symptoms, drug dosage, and response data. Its primary goal is to identify potential side effects of medications by highlighting or flagging relevant data items (Khurana et al., 2023).

Another NLP system, developed at Columbia University in New York, is called MEDLEE (MEDical Language Extraction and Encoding System). MEDLEE is designed to identify clinical information from narrative reports and convert textual information into a structured representation (IHME, 2023).

Other significant applications that NLP has had in the field of healthcare include:

- Medical Information Extraction from EHR systems. This usecase has been adopted for identification of patient phenotype cohorts, smoking status extraction, genome-wide association studies, extraction of adverse drug events, detection of medication discrepancies, temporal relation discovery, risk stratification, and risk factor identification from EHR data (Khanbhai et al., 2021; Wang et al., 2018).

- Sentiment Analysis to evaluate patient perception. In (Greaves et al., 2013) used sentiment analysis techniques to categorize online free-text comments by patients as either positive or negative descriptions of their health care.

  Subsequently, (van Buchem et al., 2022) developed an AI-based tool, consisting of a questionnaire, NLP pipeline, topic modeling and visualization to enable physicians to evaluate and prioritize patient experiences without being confined to the limited answer options of closed-ended questions.

- Automatic Translation and Summary Generation of medical information in various languages. Machine Translation is now commonly used by translation vendors and in business applications, where it is either used as the sole system or in combination with other systems or human post-editing. I healthcare, it might be used to clarify patient histories, review a clinical diagnosis, or restate the recommended treatment plan and follow-up to facilitate comprehension. However, Machine Learning in health communication is in an initial step to be followed by human correction (Dew et al., 2018; Randhawa et al., 2013).

- Early Detection of Cognitive Impairment through Language Analysis. A study was conducted by (Clarke et al., 2021) where a high-dimensional set of linguistic features were automatically extracted from each patient transcript and used to train Support Vector Machines to classify groups, obtaining moderate results.

Referring specifically to the topic of our study, although the number of publications is not large yet, an increasing trend is noticed on the number of works motivated by NLP as a Quality of life evaluation tool for Chronic patients.

By developing a tool incorporating NLP, it becomes possible to assess the impact of chronic conditions on patients' health-related quality of life, encompassing their physical and emotional well-being, psychosocial adaptation, and overall satisfaction, whether derived from medical records or self-reports

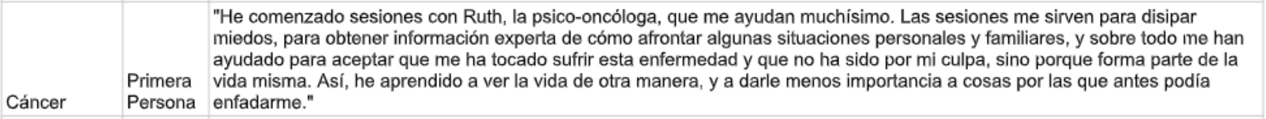
Moreover, this tool holds the potential to significantly reduce the time and resources required for text labeling and analysis (Forsyth et al., 2018). Such efficiency gains would benefit both the socio-health system and healthcare professionals, resulting in time and cost savings. Furthermore, the scalability of this tool extends to other domains, including social and health research, where incorporating health outcomes and patients' experiences in healthcare processes is essential.

3. Methodology

As mentioned in previous chapters, this study will focus on testimonies. Testimonies from patients with chronic diseases will serve as the foundation from which we will aim to extract knowledge.

Regarding the collection of testimonies at our disposal, it has been obtained through an Internet search. The objective of the search was to gather diverse testimonies from multiple sources, representing patients with various diseases. The main sources for the testimonies were the Josep Carreras Foundation and A Breath of Hope Lung Foundation.

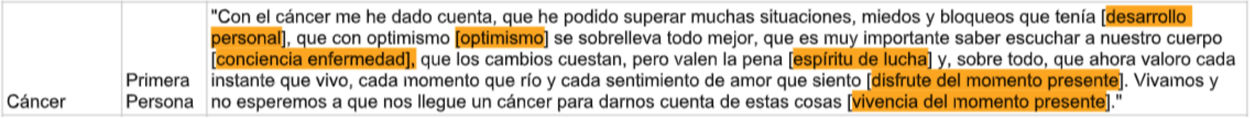
All the testimonies share that they are in Spanish, mostly in Castilian Spanish, but in some cases also in Latin American Spanish. Among them, a variable feature is the narrator: approximately half of the cases were narrated in first person (by the patients themselves), while the other half were narrated in third person (family members, partners, acquaintances, etc.). An example of a transcription of a testimony is provided below:



Cáncer	Primera Persona	"He comenzado sesiones con Ruth, la psico-oncóloga, que me ayudan muchísimo. Las sesiones me sirven para disipar miedos, para obtener información experta de cómo afrontar algunas situaciones personales y familiares, y sobre todo me han ayudado para aceptar que me ha tocado sufrir esta enfermedad y que no ha sido por mi culpa, sino porque forma parte de la vida misma. Así, he aprendido a ver la vida de otra manera, y a darle menos importancia a cosas por las que antes podía enfadarme."

Figure 3.1: Testimony example

After obtaining a (moderate) battery of testimonies, they were labeled by a healthcare professional from the ProHealth research group at the Valencia International University (VIU). An example of such labeling can be seen in the following figure:



Cáncer	Primera Persona	"Con el cáncer me he dado cuenta, que he podido superar muchas situaciones, miedos y bloqueos que tenía [desarrollo personal], que con optimismo [optimismo] se sobrelleva todo mejor, que es muy importante saber escuchar a nuestro cuerpo [conciencia enfermedad], que los cambios cuestan, pero valen la pena [espíritu de lucha] y, sobre todo, que ahora valoro cada instante que vivo, cada momento que río y cada sentimiento de amor que siento [disfrute del momento presente]. Vivamos y no esperemos a que nos llegue un cáncer para darnos cuenta de estas cosas [vivencia del momento presente]."

Figure 3.2: Labelling example

Throughout the text, we can observe that factors or indicators of Quality of Life, denoted in orange and within brackets, are highlighted, referencing the previous sentence. These examples consist of very short testimonies for representation purposes, however, in our collection we have texts of varying lengths, ranging from less than 50 words to over 500.

Nevertheless, due to time constraints and the effort required to manually annotate texts like the ones mentioned above, combined with the high value of a healthcare professional's time, the number of testimonies that have been labeled is limited. Thus, we are left with a quite small dataset consisting of 50 testimonies to work with in model development. We will separate 5 of these testimonies to test the developed models on unseen data.

As mentioned in the introduction of this document, the project we are undertaking is part of an initiative that will lead to the publication of a research article. This Master's Thesis only represents an initial approach to a much more complex project led by the ProHealth research group.

The large-scale project will aim to identify indicators within testimonials from a fixed library of quality of life factors. Given the highly challenging nature of this problem combined with the limited availability of labeled testimonies, this study will focus on alternative approaches or partial solutions that allow for an initial exploration of the problem and its context.

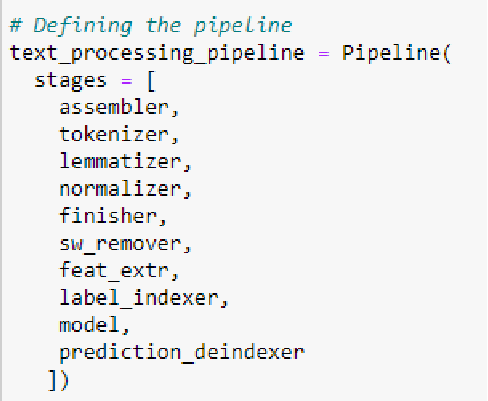
SparkNLP

To address the problem at hand, we used dedicated software for Natural Language Processing, specifically the SparkNLP library. SparkNLP is an open-source library developed by the John Snow Labs team, who defines itself as an award-winning AI company that specializes in assisting healthcare and life science organizations in leveraging AI. They provide high-compliance AI software, models, and data.

This library is specifically designed to operate within the distributed processing system environment of Apache Spark (Zaharia et al., 2010), which is also open-source. Spark is a distributed processing framework known for introducing two in-memory data structures called Resilient Distributed Datasets (RDD) and Dataframes. Moreover, it has shown superior performance compared to Hadoop, especially in iterative Machine Learning jobs (Zaharia et al., 2010).

SparkNLP offers various language preprocessing functionalities, including tokenization, lemmatization, normalization, and stopword removal. It is especially useful when intending to efficiently and distributedly process large volumes of data. It provides a comprehensive library of pre-trained models for different types of tasks. These models can be used directly or be fine-tuned to adapt them to specific problems.

The utilization workflow of this library is based on the creation and execution of pipelines. In the context of Data Science, a pipeline is an ordered sequence of steps performed to process, transform, and train models on data. It provides a structured and automated approach to organize different stages of the workflow. In an NLP project like the one at hand, a pipeline may include various transformations for data preprocessing, feature extraction, and the training of a prediction model using these features.



```
# Defining the pipeline
text_processing_pipeline = Pipeline(
  stages = [
    assembler,
    tokenizer,
    lemmatizer,
    normalizer,
    finisher,
    sw_remover,
    feat_extr,
    label_indexer,
    model,
    prediction_deindexer
   ])
```

Figure 3.3: NLP Pipeline Example

In the previous example, we observe a list of different modules, one after another, each responsible for one of the mentioned process tasks in Section 2, such as text preprocessing (tokenizer, lemmatizer, normalizer, sw_remover), feature extraction (feat_extr), and the model that will utilize the aforementioned features. Prior to defining the pipeline, each of these modules must be individually defined, and their respective parameters need to be set.

Cloud Computing: AWS

In order to accomplish everything mentioned thus far in this section, the computing power provided by the available commodity computer has proven to be insufficient. Therefore, we have required support from Cloud Computing, specifically the services of Amazon Web Services (AWS).

Cloud Computing refers to both the applications delivered as services over the Internet, and also the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing (Armbrust et al., 2009). This is the case of Amazon web Services (AWS), which, to this date, stands as the largest and most well-known public cloud provider.

AWS offers a wide range of cloud services, including compute power, storage, databases, networking, and many others. It enables organizations to scale their infrastructure and resources on-demand, paying only for what they use. This, added to its extensive global presence and robust feature set and the rich ecosystem of tools it provides, has allowed AWS to become a go-to choice for businesses of all sizes looking to leverage the benefits of Cloud Computing.

For the development of this study only a few of the numerous services AWS offers have been needed:

- **EC2:** It stands for Elastic Compute Cloud and is a computing service available on the AWS cloud platform. It provides on-demand access to a wide range of computing resources and offers the flexibility of a virtual environment. With EC2, users have the ability to customize their instances based on their specific requirements. This includes allocating the desired amount of RAM, ROM, and storage to efficiently handle their current tasks.

  In the SageMaker service configuration that will be explained later, we choose to launch an ml.t3.xlarge instance. This means that we launch an underlying EC2 instance with specific characteristics and configurations to support the Machine Learning functionalities in SageMaker. These types of instances offer a combination of CPU and memory resources suitable for Machine Learning workloads. Additionally, they provide GPU acceleration resources for more intensive tasks such as those related to Deep Learning.

- **S3:** Amazon defines S3 (Simple Storage Service) as an object storage service that offers industry-leading scalability, data availability, security, and performance. It is used by customers of all sizes and industries. Amazon S3 provides management features to optimize, organize, and configure data access to meet business requirements.

  We do not require a large amount of storage or specific data access configurations. Therefore, for this project, it will suffice to launch a generic S3 instance and create a bucket to store the various files generated. A bucket is simply the name given to containers for objects stored in Amazon S3. They can store any number of objects, and up to 100 buckets can be created per AWS account.

- **SageMaker:** It is a fully-managed AWS platform that enables the creation, training, and deployment of Machine Learning models. It covers the entire workflow, including data labeling and preparation, algorithm selection, training, execution, and deployment.

  Within the SageMaker platform, there are numerous services, among which we will benefit from Amazon SageMaker Studio Notebook Instances. This allows us to launch fully-managed, standalone Jupyter Notebook instances in the Amazon SageMaker console, with the flexibility to choose from a wide range of cloud computing resources.

In the following sections, we will explain the approaches we have taken to the problem of identifying indicators of Quality of Life in testimonies and how we have attempted to address them.

Among the analyzed testimonies, a total of 287 indicators of Quality of Life have been identified, distributed across 657 occurrences. Some indicators are repeated numerous times, while others may appear only once.

It would be naive to try to identify indicators in the text from a library of over 250 possibilities, considering that we only have 50 testimonies, 5 of which are reserved to test the developed models on unseen data. That is why we have decided to reduce the number of possible labels based on the quality of life dimensions analyzed by the SF-36 Quality of Life Assessment Questionnaire (see Chapter 2).

In this sense, we have rigorously mapped the indicators to each of the 8 dimensions of the SF-36 questionnaire. In addition to these 8 dimensions, we have found it necessary to add a ninth label to assign to indicators related to interactions with medical personnel, as they are quite frequent in this type of testimonies. Finally, another very minor class has been added to capture sporadic indicators that do not fit into any of the other 9 classes. The mapping is as is shown in the following diagram:

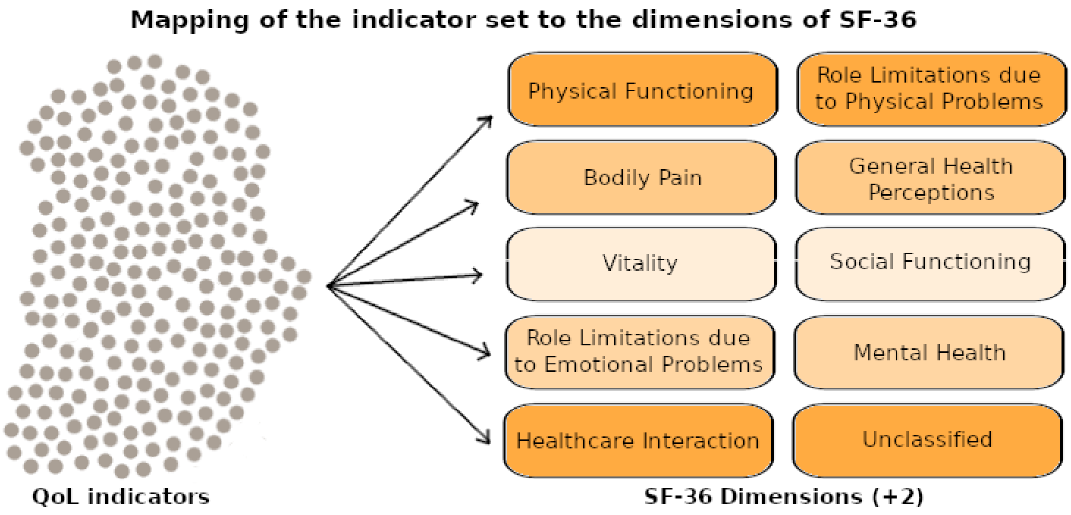

Mapping of the indicator set to the dimensions of SF-36
Physical Functioning
Role Limitations due to Physical Problems
Bodily Pain
General Health Perceptions
Vitality
Social Functioning
Role Limitations due to Emotional Problems
Mental Health
Healthcare Interaction
Unclassified
QoL indicators
SF-36 Dimensions (+2)

Figure 3.4: Mapping of the indicator set to the dimensions of SF-36

Let's now proceed with a brief exploratory analysis of the dataset we will be working with, following the previously mentioned transformation.

We begin by examining a pie chart that illustrates the fraction of indicators classified within each dimension of the SF-36 questionnaire.

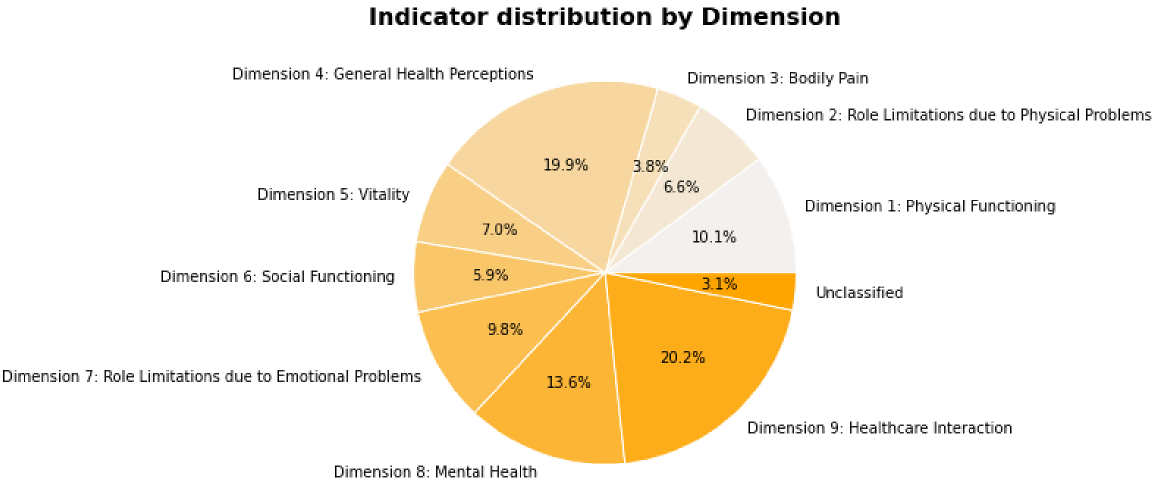


Figure 3.5: Indicator distribution by Dimension

We observe that there are 2 dimensions that stand out for gathering a large amount of different indicators, specifically Healthcare Interaction and General Health Perceptions. This simply indicates that we find these dimensions in the patients' testimonials expressed through a greater variety of indicators.

It is important not to confuse the above graph with the frequencies of occurrence for each dimension. This can be seen in the following bar graph, where each bar represents the number of times indicators of each dimension have been identified in the whole dataset.

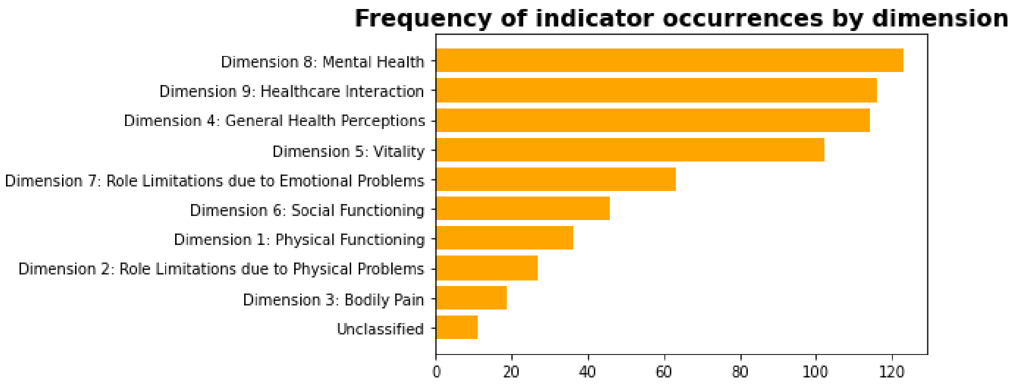


Figure 3.6: Frequency of Indicator Ocurrences

Thus, we can observe that the 4 dimensions that appear most frequently in the patients' testimonies are, in this order, Mental Health, Healthcare Interaction, General Health Perceptions, and Vitality.

On the other end of the spectrum, we have the minority classes of Unclassified and Bodily Pain, with the former simply representing indicators that could not be classified into any of the other 9 classes. Consequently, we can conclude that bodily pain is a topic that is not commonly mentioned in the testimonies of our patient sample.

Now, we will explain the two approaches we have taken, where we simplify the problem to a text classification task with two different levels of complexity.

3.1 First Approach: Multiclass Text Classification

- **Multiclass Text Classification problem**: As we have seen in Section 2, one of typical tasks in the field of Natural Language Processing is text classification into two or more distinct classes.

  In this case, our goal is to classify each sentence present in the testimonials based on the dimension of the indicator it contains, if any. We will assess the difference in performance achieved using more traditional Machine Learning-based models versus more innovative Deep Learning-based models.

  The first challenge we encounter in addressing this problem, which could significantly impact the performance of the fitted models, is the severe class imbalance in the dataset. This is illustrated in the following graph, which depicts the number of sentences from the testimonial set labeled under each dimension using a horizontal bar chart.

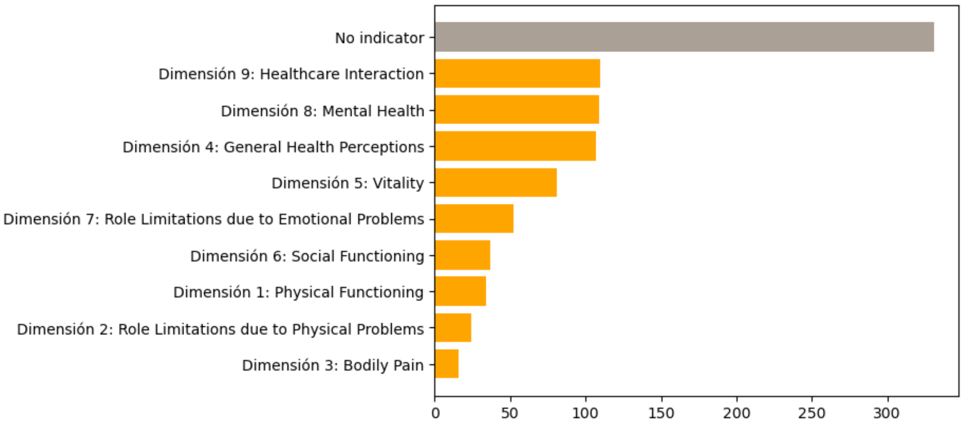


Figure 3.7: Class frquency in testimony sentences

We quickly identify that a significant portion of the sentences comprising the testimonials are not useful for analyzing patients' quality of life since they do not contain any indicators. To address this issue, we attempted to resolve it through a process known as *subsampling*. This term refers to the process of randomly selecting a sample of data from a larger set. In this case, we randomly selected 100 sentences that do not contain any indicators, resulting in the following class distribution:

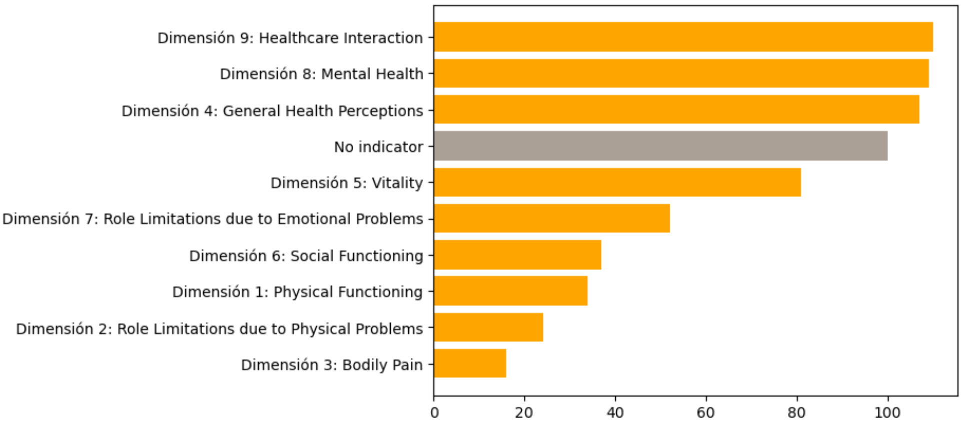


Figure 3.8: Class frquency in testimony sentences (subsampling)

We observe that, in this way, the composition of the dataset is much more balanced, which will positively affect the performance of the models we fit.

Regarding the problem-solving process, both Machine Learning-based and Deep Learning-based methods follow a similar text preprocessing pipeline. This pipeline typically includes the following Spark NLP modules:

- *DocumentAssembler()*: Transforms each sentence into a SparkNLP document.
- *Tokenizer()*: Tokenizes, i.e., converts each sentence into a series of words.
- *LemmatizerModel.pretrained()*: Reduces words (tokens) to their base form or lemma.
- *Normalizer()*: Cleans the text by removing special characters, converting to lowercase, removing spaces, etc.
- *StopWordsRemover()*: Eliminates stopwords from the text, i.e., words that do not contribute any meaningful information.
- *StringIndexer()*: Converts the labels from strings to integers so that they can be processed by the models.
- **Feature Extraction**: This module varies depending on the method and the type of model we want to fit. We address this point separately in each method.

- *IndexToString()*: Reverses the transformation of labels, converting integer values back to their original string values.

Let us now examine the specifics of each of the methods employed to address the problem and the distinct models that will be fitted for each of them.

- **ML-based Method:** What distinguishes this method is the fact that we do not apply Deep Learning techniques for text feature extraction or classification models.

  For feature extraction, we use Bag of Words and TF-IDF, while for classification based on these features, we utilize three commonly used Machine Learning models: Naïve Bayes, Logistic Regression, and Random Forest Classifier. To bridge any possible knowledge gap, we will briefly explain the intuition behind these Machine Learning models:

  Naïve Bayes is a Machine Learning model that applies Bayes' Theorem for probabilistic classifications, assuming conditional independence among variables or features. On the other hand, Logistic Regression uses a sigmoid function to model the relationship between variables and the probability of belonging to a certain class. Finally, Random Forest combines multiple weak decision trees to create a more robust model with improved generalization and reduced overfitting.
- **DL-based Method:** In contrast to the previous method, we apply Deep Learning techniques for both text feature extraction and classification.

  For text feature extraction, we utilize three different embeddings: GloVe, BERT, and BIO-roBERTa. The first two have been extensively explained in Chapter 2:Theoretical Framework. As for BIO-roBERTa, it is a version of the pre-trained roBERTa model but basing its training on a large corpus of medical and biological language primarily sourced from the PubMed database.

  Additionally, RoBERTa (Robustly Optimized BERT Approach) is a language model based on the Transformer architecture, developed by Meta. It is built upon the success of BERT and focuses on enhancing performance and robustness (**liu**).

In both methods, a hyperparameter tuning is performed to obtain the best possible model. In the case of ML-based models, the specific hyperparameters of each model are adjusted, such as ($maxIter$, $regParam$) in Logistic Regression or ($numTrees$, $maxDepth$) in Random Forest. On the other hand, for DL-based models, the same parameters are always adjusted: ($batch\_size$, $dropout$, $learning\_rate$).

3.2 Second Approach: Binary Text Classification

- **Binary Classification Problem:** The classification problem mentioned in the previous point significantly reduces the complexity of the initial problem. However, the number of classes remains quite high compared to the size of the data. Let us recall that we only have 50 labeled testimonies.

  With that being said, we have further simplified the problem, reducing its complexity even more. In this case, we have limited ourselves to labeling the sentences that compose the testimonies based on whether or not they contain a Quality of Life indicator, regardless of the dimension it belongs to.

  This way, we transform the problem into a binary classification task, which can be more manageable given the characteristics and resources of this project.

  The procedure carried out to solve this problem is analogous to the multi-class case, only restricting the number of output classes to 2. One difference we encounter compared to the previous point is that, in this case, for the Machine Learning-based method, we have been able to add 2 additional types of models, also varying their respective parameters. We are referring to Linear Support Vector Machines and Gradient Boosting Trees.

  On one hand, Linear Support Vector Machines utilize SVMs with a linear kernel to optimally separate classes in a feature space. On the other hand, Gradient Boosting Trees combine multiple decision trees through the gradient boosting technique, allowing for the capture of non-linear relationships and handling complex features.

3.3 Metrics

We shall select a set of metrics to assess the performance of the various Machine Learning models we train to accomplish the task at hand. It is common to evaluate the correctness of a classification task by computing the number of correctly identified class examples (true positives), the number of correctly identified examples that do not belong to the class (true negatives), as well as the examples that were incorrectly assigned to the class (false positives) or were not recognized as class examples (false negatives) [40]. These four counts form a confusion matrix, of which we can observe an example in the following figure:

		Predicted label	
		1	0
Real label	1	*tp*	*fn*
	0	*fp*	*tn*

This can be mathematically expressed as a matrix and a variety of metrics are derived from it:

$$CM = \begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

$$Precision\ (P) = \frac{tp}{tp + fp}$$

$$Recall(R) \ = \frac{tp}{tp + fn}$$

$$F1 - score(F1) \ = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

- **Precision:** It is calculated by dividing the number of correctly classified positive examples by the total number of examples labeled as positive by the system (Sokolova & Lapalme, 2009). It tells us how many of the predicted positive results were actually correct.
- **Recall:** It is determined by dividing the number of correctly classified positive examples by the total number of positive examples in the dataset (Sokolova & Lapalme, 2009). It indicates how many of the actual positive instances were successfully captured by the model.
- **F1-score:** It is a combined metric that takes into account both precision and recall (Sokolova & Lapalme, 2009). It provides a single value that balances the trade-off between precision and recall.

The F1-Score is a metric that combines precision and recall, using the harmonic mean to keep a balance between the two ratios. It provides a single measure that is more sensitive to low values in both of these metrics, ensuring a balanced evaluation of classification models.

Due to its ability to balance precision and recall, the F1-Score has gained importance as a reliable metric for evaluating classification models. By emphasizing the importance of both, it offers a comprehensive assessment of a model's performance.

As a result, it has become our preferred metric for assessing the effectiveness of classification models.

4. Results

In this chapter, we will analyze the results obtained by each method in each of the approaches we have taken towards solving the problem.

4.1 First Approach: Results of the Muliclass Classification Problem

Note that the initial approach we have taken to tackle the problem was to reduce it to a classification task with a total of 10 different classes.

Using the metrics explained in the previous section, primarily the $F1\_score$, we will analyze the results obtained for each type of model we have attempted to fit, including their different versions obtained in the hyperparameter tuning study.

With the objective of understanding of how well the models of each type have performed overall in solving the current problem, we generate a bar graph that illustrates their average $F1\_score$. It is presented below:

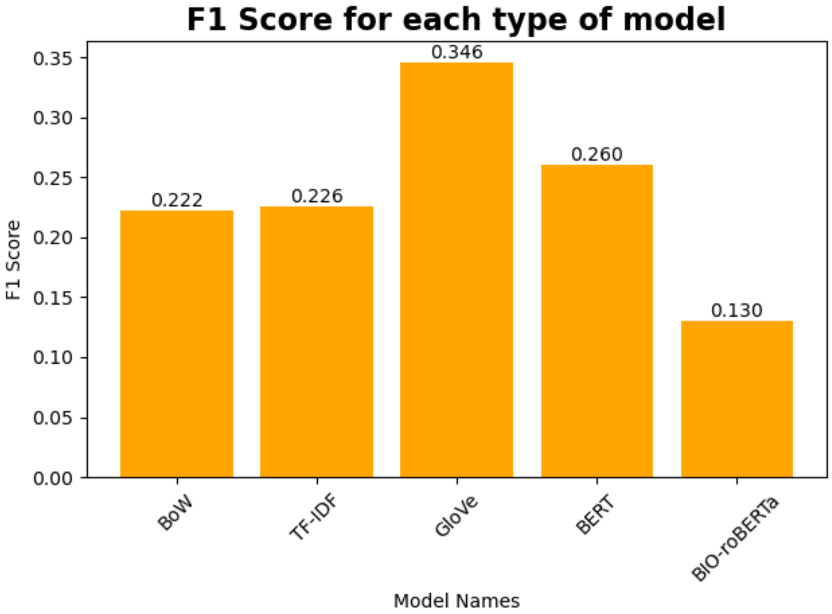


Figure 4.1: F1 Score for each type of model

We can observe that the average $F1\_score$ of DL-based models is higher than that of ML-based models, except for those utilizing BIO-RoBERTa embeddings, which in this case have the lowest $F1\_score$. We can conclude that, in general terms, the models based on GloVe embeddings demonstrate better performance for this problem. However, we notice that the $F1\_score$ values are not particularly high.

That being said, it is not of particular interest for us to know which models perform better in general, but instead we seek the best model for this specific task. For this reason, we select the top 5 models of each type and repeat the previous graph. This approach will allow us to identify the model types that best address our classification problem.

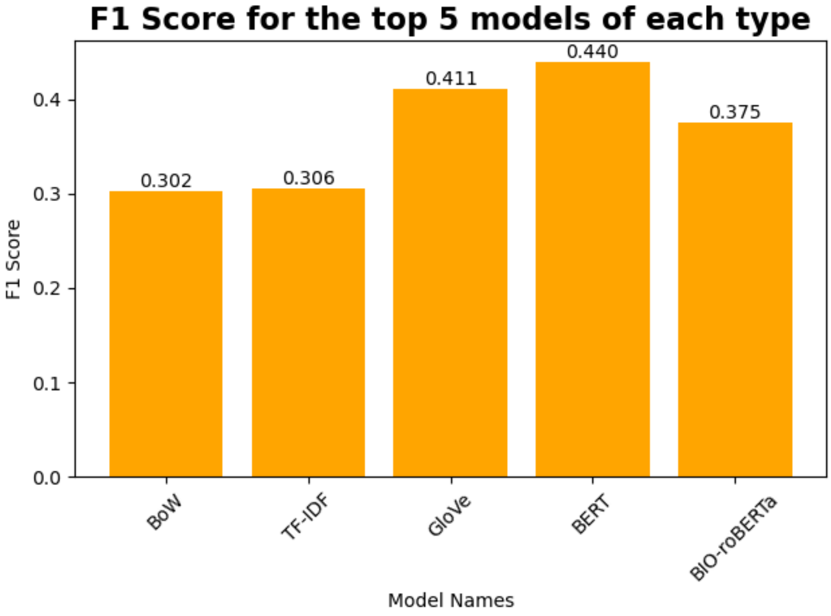


Figure 4.2: F1 Score for the top 5 models of each type

We observe that, as expected, the values of $F1\_score$ are slightly higher in this case. The performance of DL-based models is significantly better than ML-based models. An interesting finding is that, unlike the previous case, it is the BERT-based models that appear to be more effective.

In addition, there is a noticeable difference in the BERT and BIO-RoBERTa columns compared to the previous graph, indicating that different combinations of hyperparameters have produced diverse and inefficient results. However, in the best cases, they outperform the ML-based models.

To complement the aforementioned conclusions, we have filtered all DL-based models based on whether they have an $F1\_score \geq 0.4$, and we present a graph depicting the distribution of embedding types among the models that meet this condition.

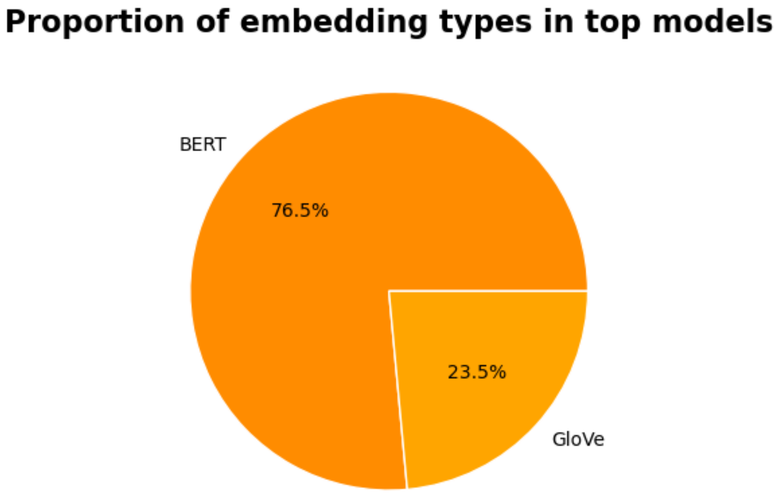


Figure 4.3: Proportion of embedding types in top models

We observe that models with an $F1\_score$ higher than 0.4 are solely based on BERT and GloVe embeddings, with 76.5% of them corresponding to BERT and the remaining 23.5% to GloVe.

The visualization of this graph, along with the previous one, allows us to conclude that although, in general terms, GloVe embeddings obtain the best results across different combinations of hyperparameters, this search for optimal hyperparameters enables us to find specific models based on BERT embeddings that outperform GloVe.

The fact that there are no models based on BIO-RoBERTa embeddings among the top 10 indicates that, despite not obtaining the worst results, it is not the ideal option for solving this problem.

Lastly, let us examine a table presenting information about the best models we have fine-tuned for the multiclass classification task and select the optimal one:

Embeddings	Batch	Dropout	Learning rate	Epochs	Train F1	Test F1
BERT	4	0.1	0.003	75	0.600	0.449
BERT	8	0.1	0.001	75	0.596	0.443
BERT	16	0.1	0.003	75	0.588	0.442
BERT	16	0.2	0.003	75	0.580	0.441
GloVe	4	0.1	0.003	75	0.551	0.426
BERT	16	0.1	0.001	75	0.571	0.424
GloVe	2	0.2	0.003	75	0.532	0.424
BERT	8	0.1	0.003	75	0.665	0.420
BERT	2	0.2	0.001	75	0.597	0.412
BERT	2	0.1	0.001	75	0.698	0.411

Table 4.1: 10 mejores modelos obtenidos de la búsqueda de hiperparámetros

As mentioned previously, we observe that the best models are based on BERT embeddings and have different combinations of hyperparameters. Just 2 GloVe models are able to get a spot in the 10 best models but the top 4 are still based on BERT embeddings.

Therefore, we choose the following model as the best one to solve the proposed multiclass classification problem:

DL-based method, utilizing BERT embeddings for text feature extraction, trained for 75 epochs, with a batch size of 4, learning rate of 0.003, and dropout of 0.1.

Evaluation on Unseen Data

Remember that out of the 50 labeled testimonies we dispose of, we have set aside 5 (10%) for testing the optimal model on unseen data. Therefore, we will classify the sentences comprising these 5 testimonies based on the dimension to which the indicator they contain belongs, if any.

The classification will result in a confusion matrix, where we can observe the original classes extracted from the labeled data compared to the classes predicted by the model. This allows us to examine the model's behavior, whether it performs better on certain classes than others and any other biases.

Let us now proceed to examine the confusion matrix below:

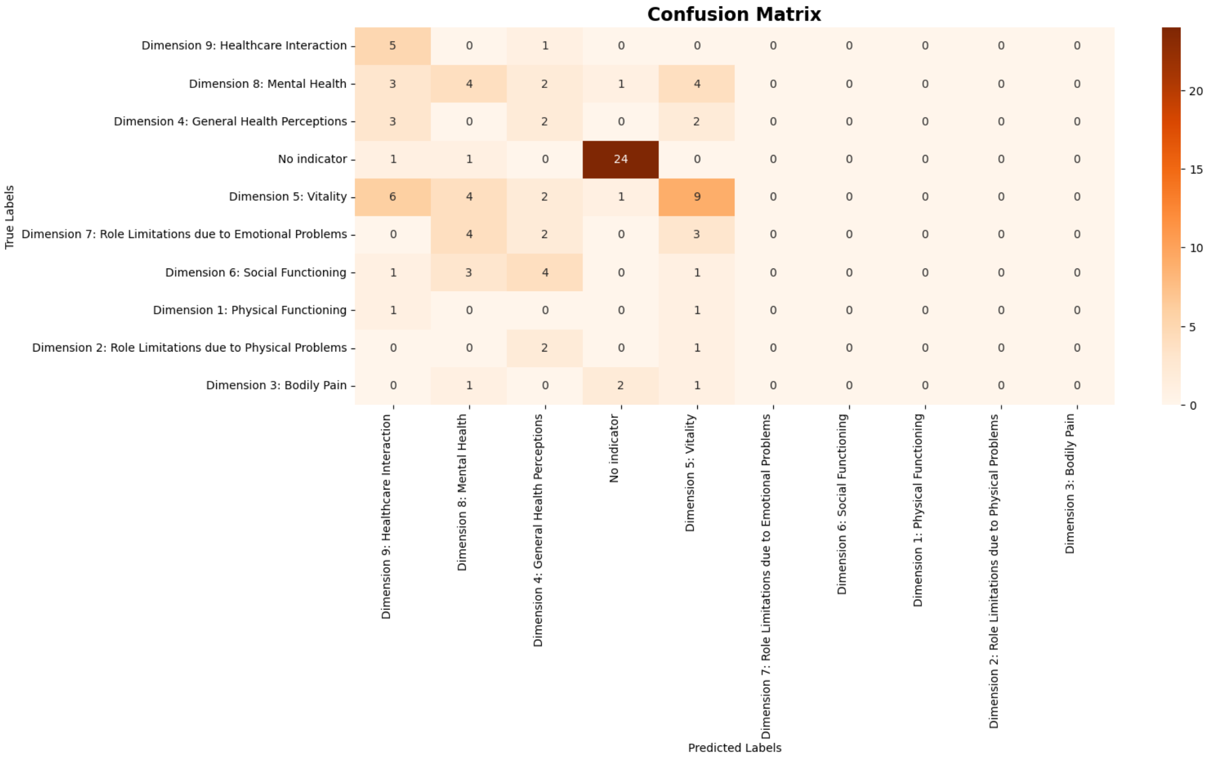


Figure 4.4: Confusion Matrix for best Multiclass Classification Model

It is evident that the results are quite poor, obtaining a moderately good performance in certain classes such as the *No indicator* and *Vitality* classes.

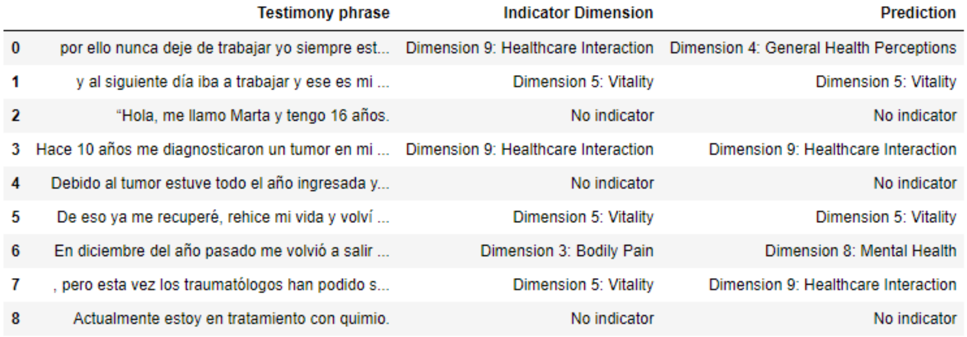
It is apparent that the model has only attempted to classify the sentences into 5 out of the 10 possible classes. Not coincidentally, these 5 classes correspond to the 5 most represented classes in the dataset. This highlights the effect of the scarcity of labeled data, especially in certain dimensions.

The aforementioned confusion matrix translates into the following metrics:

- **Precision: 0.363**
- **Recall: 0.431**
- **F1-Score: 0.385**

Despite being the chosen model, its metrics indicate that the DL-based model is not achieving accurate and efficient classification. An $F1\_score$ of 0.385 on unseen data suggests that the model struggles to generalize and correctly discriminate between different classes. As we have seen, the existing class imbalance and lack of labeled data for model training negatively affect its performance, biasing it towards certain classes.

Let's now examine some examples of the model's classifications compared to the actual labels:



	Testimony phrase	Indicator Dimension	Prediction
0	por ello nunca deje de trabajar yo siempre est...	Dimension 9: Healthcare Interaction	Dimension 4: General Health Perceptions
1	y al siguiente día iba a trabajar y ese es mi ...	Dimension 5: Vitality	Dimension 5: Vitality
2	“Hola, me llamo Marta y tengo 16 años.	No indicator	No indicator
3	Hace 10 años me diagnosticaron un tumor en mi ...	Dimension 9: Healthcare Interaction	Dimension 9: Healthcare Interaction
4	Debido al tumor estuve todo el año ingresada y...	No indicator	No indicator
5	De eso ya me recuperé, rehice mi vida y volví ...	Dimension 5: Vitality	Dimension 5: Vitality
6	En diciembre del año pasado me volvió a salir ...	Dimension 3: Bodily Pain	Dimension 8: Mental Health
7	, pero esta vez los traumatólogos han podido s...	Dimension 5: Vitality	Dimension 9: Healthcare Interaction
8	Actualmente estoy en tratamiento con quimio.	No indicator	No indicator

Figure 4.5: Example Predictions for best Multiclass Classification Model

The model performs moderately well for certain classes, but overall, the problem has proven to be too complex given the available resources.

4.2 Second Approach: Results of the Binary Classification Problem

Considering the (already confirmed) possibility that the previous approach was still too complex for the available labeled data and resources, a second approach involved further reducing the number of classes, resulting in a binary classification problem. Our goal is to determine, for each sentence in a testimony, whether it contains a Quality of Life indicator or not.

The development of the models and their training is essentially equivalent to the previous case, and therefore, the obtained results follow the same structure. We begin by examining a bar chart depicting the average $F1\_score$ of the different fitted models for each of the 5 types (bare in mind: BoW and TF-IDF corresponding to ML-based methods, and GloVe, BERT, and BIO-roBERTa corresponding to DL-based methods).

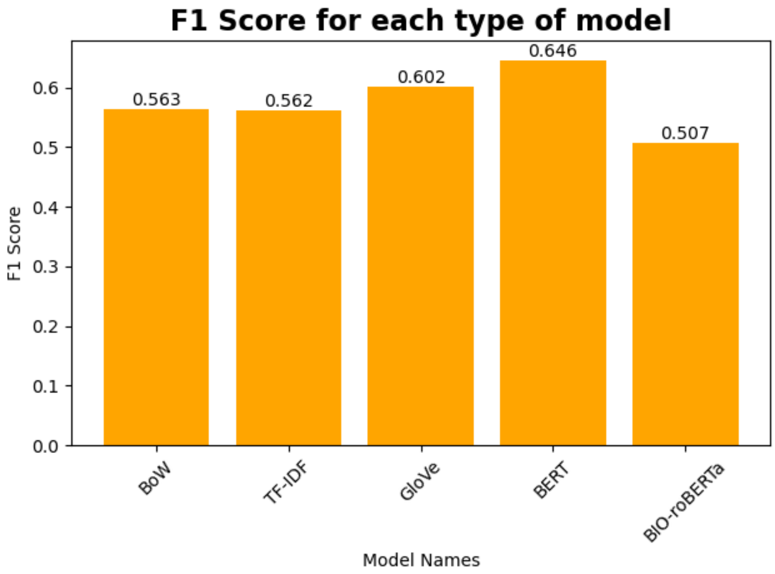


Figure 4.6: F1 Score for each type of model

We can observe that, in this case, the difference in the average $F1\_score$ is not as significant among the different types of models considered. We appreciate that the $F1\_score$ values, ranging around 0.6, are now significantly higher than in the previous problem. Following the same approach as in the previous section, let's examine the comparison not among all models, but only among the top 5 of each type:

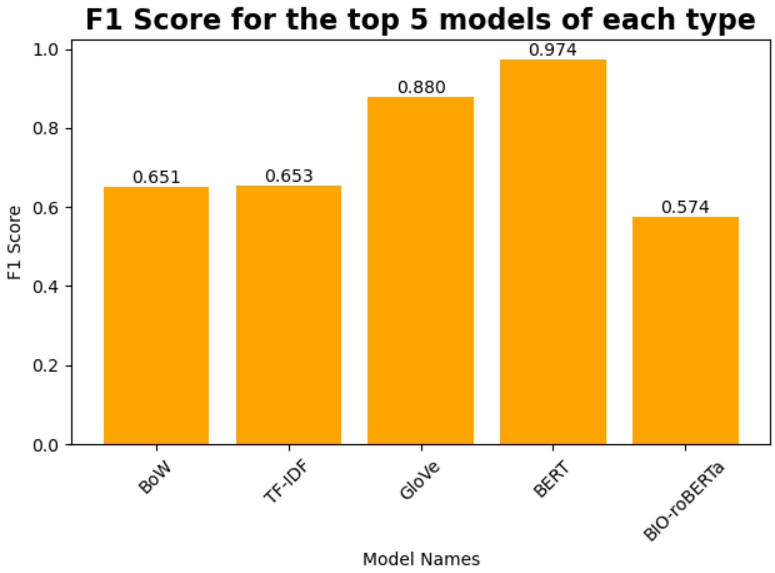


Figure 4.7: F1 Score for the top 5 models of each type

When considering only the top-performing models, the gap between ML-based methods and DL-based methods becomes significantly larger, with the former achieving $F1\_score$ values close to 0.65, while the latter reaching values up to approximately 0.97. The only exception to this rule is observed in the models based on BIO-RoBERTa embeddings, which are performing below expectations considering their complexity and specific training on healthcare texts.

Although the first graph might suggest no major overall difference between using one type of model or the other, focusing on the top-performing models indeed reveals a significant distinction.

This finding also indicates that the performance variability among DL-based methods is greater compared to ML-based methods. It is worth highlighting that, similar to the previous case, the BERT-based models seem to exhibit higher performance in solving the problem.

Let's proceed with the pie chart, this time filtering all DL-based models based on whether they meet the condition $F1\_score \geq 0.85$. Let's see if, as in the previous case, there are dominant model types.

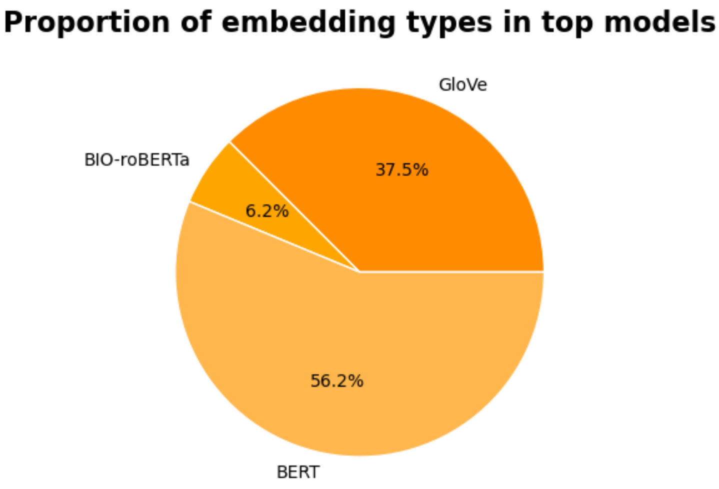


Figure 4.8: Proportion of embedding types in top models

In this case, the top-performing models are slightly more distributed among the different embeddings used, although we still find that those based on BERT remain predominant.

A notable fact is that among the top models, there is one based on BIO-RoBERTa embeddings. This indicates that it is not that this feature extraction method is inefficient, but rather that the chosen combinations of hyperparameters have not favored the performance of this type of model. However, we see that these embeddings have the potential to generate a more than acceptable performing model.

Similar to the multiclass classification problem, we present below the list of the best models obtained through hyperparameter tuning, and we will select the best one among them.

Embeddings	**Batch**	**Dropout**	**Learning rate**	**Epochs**	**Train F1**	**Test F1**
BERT	64	0.000	0.001	20	0.989	0.978
BERT	32	0.200	0.001	20	0.993	0.978
BERT	64	0.200	0.001	20	0.986	0.972
BERT	32	0.400	0.001	20	0.986	0.972
BERT	64	0.400	0.001	20	0.986	0.972
BERT	32	0.000	0.001	20	0.992	0.972
BIO-RoBERTa	32	0.200	0.001	20	0.977	0.944
GloVe	16	0.000	0.001	20	0.971	0.906
GloVe	64	0.000	0.001	20	0.968	0.888
GloVe	16	0.400	0.001	20	0.924	0.872

Table 4.2: 10 mejores modelos obtenidos de la búsqueda de hiperparámetros

Once again, we observe a strong dominance of BERT-based embedding models, followed by the top-performing GloVe models. As mentioned before, one model stands out: BIO-RoBERTa, achieved through a fortunate combination of hyperparameters.

The majority of BERT models leading the table all have F1-scores above 0.97, indicating exceptionally good performance with limited room for further improvement. We select the optimal model for the binary text classification task:

DL-based method utilizing BERT embeddings for feature extraction, trained for 20 epochs with a batch size of 64, learning rate of 0.001, and no dropout.

Evaluation on Unseen Data

Similar to the previous problem, we test the selected model on the 5 unseen testimonials. We obtain a confusion matrix that can be observed below:

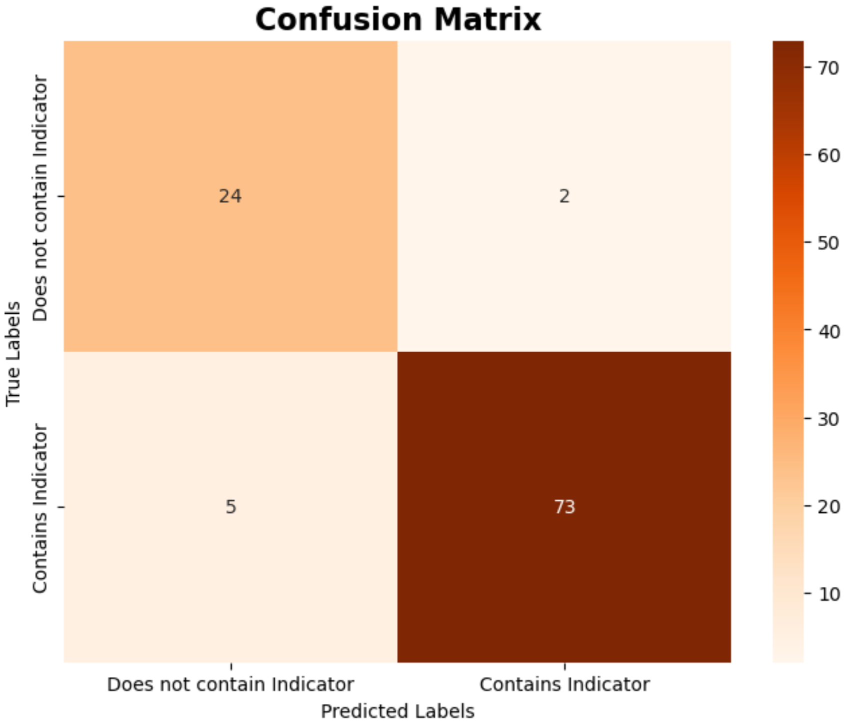


Figure 4.9: Confusion Matrix - Binary Classificaiton

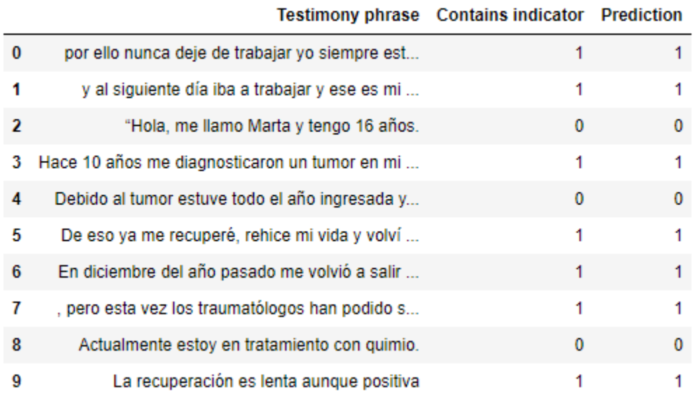
It is evident that the model has no issues in differentiating sentences based on whether they contain Quality of Life indicators or not. It correctly classifies 73 out of the 78 sentences that contained indicators and accurately discards 24 out of the 26 sentences that did not.

From the previous confusion matrix, the following metrics are obtained:

- **Precision: 0.937**
- **Recall: 0.933**
- **F1-Score: 0.934**

The fact that our DL-based learning model achieves an F1-score of 0.934 on unseen data demosntrates its excellent performance in text classification for the two different classes. Such a high F1-score confirms that the model achieves a balance between precision and recall, meaning it can correctly classify the majority of positive and negative observations in the set of sentences extracted from testimonials.

This model is hard to improve upon and can be part of a tool that significantly streamlines the work of medical personnel. Next, we will breefly examine some results comparing the actual label with the model's prediction.



	Testimony phrase	Contains indicator	Prediction
0	por ello nunca deje de trabajar yo siempre est...	1	1
1	y al siguiente día iba a trabajar y ese es mi ...	1	1
2	“Hola, me llamo Marta y tengo 16 años.	0	0
3	Hace 10 años me diagnosticaron un tumor en mi ...	1	1
4	Debido al tumor estuve todo el año ingresada y...	0	0
5	De eso ya me recuperé, rehice mi vida y volví ...	1	1
6	En diciembre del año pasado me volvió a salir ...	1	1
7	, pero esta vez los traumatólogos han podido s...	1	1
8	Actualmente estoy en tratamiento con quimio.	0	0
9	La recuperación es lenta aunque positiva	1	1

Figure 4.10: Example Predictions - Binary Classification

5. Conclusions

In this final chapter, to wrap up this study, we will present the conclusions, mention the limitations, and propose several ideas and directions for future work towards a potential continuation of the project.

5.1 Conclusions

- Throughout this study, a preliminary development of Artificial Intelligence models has been conducted to tackle the task of identifying Quality of Life features or indicators in testimonials from patients with chronic diseases.
    - A first approach, with a relatively high level of complexity, involved classifying the different phrases composing a testimony into one of ten classes based on the dimension of the SF-36 questionnaire to which the indicator present in the phrase belonged, if any.

      The best of the developed models obtained modest results, with an F1-score of 0.385 on unseen data. There is room for improvement, which could be completed by obtaining a more comprehensive and balanced dataset, as well as investigating data augmentation techniques.

    - A second approach, with a moderate level of complexity, aimed at classifying each phrase in a testimony based on whether it contained a quality of life indicator or not, in other words, a binary classification problem.

      In this case, much more promising results were obtained, reaching an F1-score value of 0.934 on unseen data. This model demonstrated the ability to distinguish perfectly between the two classes, maintaining exceptionally high precision and recall values.

      This model becomes a perfect candidate for integration into a potential application that could significantly facilitate the work of healthcare professionals.

    - In both approaches, the selected optimal model corresponded to the Deep Learning method, utilizing it for both text feature extraction and predictive model training.

      Specifically, BERT-based embeddings were used in both cases, leading us to conclude that this type of embeddings is most suitable for addressing the presented problem.

- The student has provided an extensive introduction to the field of Natural Language Processing, a field that is not covered in the Master's curriculum but complements it perfectly, enhancing the student's overall knowledge and skills. However, this introduction has only opened the doors to more complex levels within NLP that could be utilized to further advance and completion of this project.

- Knowledge from various subjects throughout the Master's program has been consolidated at different stages of the project. For instance, concepts from Data Mining have been applied in data acquisition and preprocessing. Also, platforms introduced in the Cloud Computing course have been used to develop models that expand upon the ones studied in the Machine Learning course.

- The student has participated in a project involving a multidisciplinary team. Such teams are common in the field of Data Science, providing the student with relevant and valuable experience for their future career.

5.2 Limitations

- The limited availability of labeled data for the study posed a significant constraint. Increasing the amount of data could have potentially resulted in better performance, particularly in the multiclass text classification problem.

- While the student license for computing on Amazon Web Services was helpful, it had certain restrictions. Using an unrestricted account would have allowed to test more hyperparameter combinations, potentially leading to more effective solutions.

- In addition to the data scarcity, the classes within the available data were imbalanced, which posed an additional challenge given the limited data.

- The choice of dimensions for classifying the indicators may not have been optimal. Several dimensions assessed by the SF-36 questionnaire are linguistically similar. Seeking a more appropriate division of indicators would enhance the project's performance.

5.3 Future Work

- Linked to one of the limitations, it would be highly beneficial for the project to carry out the labeling of a larger amount of data, which would allow for better training and consequently improved results.

- Given the lack of data, it would be appropriate to conduct a study of different available Data Augmentation techniques to enrich textual datasets.

- A highly useful practice that the project would greatly benefit from is the unification and systematization of data labeling based on the intended task. For example, specifying a fixed set of indicators prior to labeling, or labeling at the word level rather than at the sentence level. Additionally, more rigorous criteria could be established when classifying indicators into different dimensions of the SF-36 questionnaire.

- When exploring hyperparameters, beginning with a random search would reduce the search space for the subsequent grid search. This approach saves computational time that was previously spent on suboptimal hyperparameter combinations.

- An interesting direcion for future work is experimenting with more advanced Natural Language Processing techniques, such as Named Entity Recognition (NER). A project based on NER, with a rich and properly labeled dataset in IOB format, could be an ideal solution to the proposed problem.

- Instead of utilizing pre-trained Word Embeddings, an approach based on Transfer Learning could be employed, leveraging pre-trained models from reputable companies like John Snow Labs and subsequently fine-tuning them with our own dataset to adapt them to our specific problem.

Bibliography

Aaronson, N. K. (1989). Quality of life assessment in clinical trials: Methodologic issues. *Controlled clinical trials*, *10*(4), 195S–208S. https://doi.org/10.1016/0197-2456(89)90058-5

Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, *91*(6). https://doi.org/10.1186/s40537-019-0254-8

Alliance, N. (2022). *Invest to protect: Ncd financing as the foundation for healthy societies and economies.* https://ncdalliance.org/resources/invest-to-protect-ncd-financing-as-the-foundation-for-healthy-societies-and-economies

Armbrust, Michael, Fox, A., Armando, Griffith, Rean, Joseph, D, A., Katz, R., H, R., Konwinski, A., Andrew, Lee, G., Gunho, Patterson, A, D., Rabkin, Ariel, Stoica, & Matei. (2009). Above the clouds: A berkeley view of cloud computing.

Barber, E. L., Garg, R., Persenaire, C., & Simon, M. (2021). Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecologic Oncology*, *160*(1), 182–186. https://doi.org/10.1016/j.ygyno.2020.10.004

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python* (O. Media, Ed.). https://learning.oreilly.com/library/view/natural-language-processing/9780596803346/

Bonomi, A. E., Patrick, D. L., Bushnell, D. M., & Martin, M. (2000). Validation of the united states' version of the world health organization quality of life (whoqol) instrument. *Journal of Clinical Epidemiology*, *53*(1), 1–12. https://doi.org/10.1016/S0895-4356(99)00123-7

Bunge, M. (1975). What is a quality of life indicator? *Annals of internal medicine*, *2*, 65–79. https://doi.org/10.1007/BF00300471

Clarke, N., Barrick, T., & Garrard, P. (2021). A comparison of connected speech tasks for detecting early alzheimers disease and mild cognitive impairment using natural language processing and machine learning. *Frontiers in Computer Science*, *3*. https://doi.org/10.3389/fcomp.2021.634360

Cognetta-Rieke, C., & Guney, S. (2014). Analytical insights from patient narratives: The next step for better patient experience. *Journal of patient experience*, *1*(1), 20–22. https://doi.org/10.1177/237437431400100105

Deeplearning.ai. (2023). *A complete guide to natural language processing*. Retrieved May 30, 2023, from https://www.deeplearning.ai/resources/natural-language-processing/

Devins, G. M., Binik, Y. M., Hutchinson, T. A., Hollomby, D. J., Barré, P. E., & Guttmann, R. D. (1983). The emotional impact of end-stage renal disease: Importance of patients' perception of intrusiveness and control. *International Journal of Psychiatry in Medicine*, *13*(4), 327–343. https://doi.org/10.2190/5dcp-25bv-u1g9-9g7c

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchhoff, K. (2018). Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, *85*, 56–67. https://doi.org/https://doi.org/10.1016/j.jbi.2018.07.018

Edgar, T., & Manz, D. (2017). *Research methods for cyber security*.

Escrivá, J., Peyró, C., de la Iglesia Vaya, M., Montell, J., & Fabra, M. (2020). Aplicación de la inteligencia artificial con procesamiento del lenguaje natural para textos de investigación cualitativa en la relación médico-paciente con enfermedad mental mediante el uso de tecnologías móviles. *Revista de Comunicación y Salud*, *10*, 19–41. https://doi.org/10.35669/rcys.2020.10(1).19-41

EuroQol Group. (1990). Euroqol - a new facility for the measurement of health-related quality of life. *Health Policy*, *16*(3), 199–208. https://doi.org/10.1016/0168-8510(90)90421-9

Eurostat. (2017). *Final report of the expert group on quality of life indicators. 2017 edition*. Retrieved June 26, 2023, from https://ec.europa.eu/eurostat/web/products-statistical-reports/-/KS-FT-17-004

Fleishman, J., Cohen, J., Manning, W., & Kosinski, M. (2006). Using the sf-12 health status measure to improve predictions of medical expenditures. *Medical care*, *44*, I54–63. https://doi.org/10.1097/01.mlr.0000208141.02083.86

Forsyth, A. W., Barzilay, R., Hughes, K. S., Lui, D., Lorenz, K. A., Enzinger, A., Tulsky, J. A., & Lindvall, C. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. *Journal of Pain and Symptom Management*, *55*(6), 1492–1499. https://doi.org/10.1016/j.jpainsymman.2018.02.016

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: One embedding space to bind them all.

Goldberg, Y. (2017). *Neural network methods for natural language processing* (Vol. 37). Morgan & Claypool. https://doi.org/10.2200/S00762ED1V01-Y201703HLT037

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research*, *15*(11), e239. https://doi.org/10.2196/jmir.2721

Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of internal medicine*, *8*(118), 622–629. https://doi.org/10.7326/0003-4819-118-8-199304150-00009

IHME. (2023). *Gdb compare, institute for health metrics and evaluation*. Retrieved July 6, 2023, from https://vizhub.healthdata.org/gbd-compare/

INE. (2022). *Estadística de defunciones según la causa de muerte*. Retrieved July 6, 2023, from https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=resultados&idp=1254735573175

Keloth, V. K., Banda, J. M., Gurley, M., Heider, P. M., Kennedy, G., Liu, H., Liu, F., Miller, T., Natarajan, K., V Patterson, O., Peng, Y., Raja, K., Reeves, R. M., Rouhizadeh, M., Shi, J., Wang, X., Wang, Y., Wei, W.-Q., Williams, A. E., . . . Xu, H. (2023). Representing and utilizing clinical textual data for real world studies: An ohdsi approach. *Journal of Biomedical Informatics*, *142*, 104343. https://doi.org/10.1016/j.jbi.2023.104343

Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., & C., S. (2021). Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review. *BMJ Health Care Informatics*, *28*, e100262. https://doi.org/10.1136/bmjhci-2020-100262

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*, 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Klein, A., Cai, H., Weissenbacher, D., Levine, L., & Gonzalez, G. (2020). A natural language processing pipeline to advance the use of twitter data for digital epidemiology of adverse pregnancy outcomes. *Journal of Biomedical Informatics: X*, *8*, 100076. https://doi.org/10.1016/j.yjbinx.2020.100076

Lacroix, A., Jacquemet, S., Assal, J.-P., & Benroubi, M. (1995). The patients' voice: Testimonies from patients suffering from chronic disease [Proceedings of the Patient Education 2000 Congress]. *Patient Education and Counseling*, *26*(1), 293–299. https://doi.org/10.1016/0738-3991(95)00765-R

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Using the sf-12 health status measure to improve predictions of medical expenditures. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lindvall, C., Lilley, E., Zupanc, S., Chien, I., Udelsman, B., Walling, A., Cooper, Z., & Tulsky, J. (2018). Natural language processing to assess end-of-life quality indicators in cancer patients receiving palliative surgery. *Journal of Palliative Medicine*, *22*. https://doi.org/10.1089/jpm.2018.0326

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. The MIT Press. http://nlp.stanford.edu/fsnlp/

McRoy, S., & Ali, C. (2021). *Principles of natural language processing*. Susan McRoy. https://books.google.es/books?id=AZiRzgEACAAJ

Meeberg, G. A. (1993). Quality of life: A concept analysis. *Journal of Advanced Nursing*, *18*(1), 32–38. https://doi.org/10.1046/j.1365-2648.1993.18010032.x

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger, Eds.). *26*. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations, 746–751. https://aclanthology.org/N13-1090

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Randhawa, G., Ferreyra, M., Ahmed, R., Ezzat, O., & Pottie, K. (2013). Using machine translation in clinical practice. *Canadian family physician Medecin de famille canadien*, *59*(4), 382–383.

Robert, G., & Cornwell, J. (2013). Rethinking policy approaches to measuring and improving patient experience. *Journal of Health Services Research Policy*, *18*, 67. https://doi.org/10.1177/1355819612473583

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Soo, G. M. (2023). *Deep learning bible - 3. natural language processing - eng.* Retrieved June 27, 2023, from https://wikidocs.net/book/8027

Sprangers, M. A. (2002). Quality-of-life assessment in oncology. achievements and challenges. *Acta oncologica*, *41*(3), 229–237. https://doi.org/10.1080/02841860260088764

The Whoqol Group. (1998). The world health organization quality of life assessment (whoqol): Development and general psychometric properties. *Social Science Medicine*, *46*(12), 1569–1585. https://doi.org/10.1016/S0277-9536(98)00009-4

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing* (O. Media, Ed.). https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/

van Buchem, M. N., Neve, O. M., Kant, I. M. J., Steyerberg, E. W., Boosman, H., & Hensen, E. F. (2022). Analyzing patient experiences using natural language processing: Development and validation of the artificial intelligence patient reported experience measure (ai-prem). *BMC medical informatics and decision making*, *22*(1), 183. https://doi.org/10.1186/s12911-022-01923-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, *88*, 11–19. https://doi.org/https://doi.org/10.1016/j.jbi.2018.10.005

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/https://doi.org/10.1016/j.jbi.2017.11.011

Ware, J. E., & Sherbourne, C. D. (1992). The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. *Medical Care*, *30*(6), 473–483. http://www.jstor.org/stable/3765916

WHO. (2022a). *Noncommunicable diseases: Progress monitor 2022. world health organization.* Retrieved May 29, 2023, from https://apps.who.int/iris/handle/10665/353048.

WHO. (2022b). *Noncummunicable diseases. world health organization.* Retrieved May 29, 2023, from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Zaharia, M., Chowdhury, M., Franklin, M., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, *10*, 10–10.