UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Evaluating Patient Experience when using Digital Healthcare Services: surveys and Natural Language Processing-based methods

Margarida Semedo Pereira Paula Vicente

Mestrado em Engenharia Biomédica e Biofísica

Dissertação orientada por:
Doutora Ana Rita Mendes Londral
Professor Doutor Nuno Miguel de Pinto Lobo e Matela

2022

Acknowledgements

This research was made possible through the financial support provided by Fundação para a Ciência e Tecnologia (FCT) under the CardioFollow.AI project (DSAIPA/AI/0094/2020) and the Ethical Committee of Centro Hospitalar Lisboa Central (with study registration number CA2651). These approvals ensured the ethical and legal integrity of this study, safeguarding the rights and well-being of the participating individuals, who provided their informed consent before being interviewed and recorded.

I would like to express my sincere gratitude to the Faculty of Sciences of the University of Lisbon, for the opportunity to study here and for all the knowledge I acquired during these years of my degree. My appreciation also extends to my thesis supervisor, Professor Nuno Matela, for expertly guiding me and motivating me through my thesis experience.

I had the pleasure of working with wonderful professionals, which I must thank. In the first place, Dra. Ana Rita Londral, who welcomed me into her team, for her guidance during the period I worked with her. Despite being extraordinarily busy with her duties, Dra. Ana took time to hear, encourage and keep me on the correct path, allowing me to complete our study and conclude my master's. I am also grateful for meeting the members of the VOH.CoLAB team, for all the thoughtful comments and recommendations on this dissertation and for kindly integrating me in the team. Namely, I want to thank Ana Catarina Teixeira, who took the time to complete my dataset and contributed to my study results.

A special thank you to my college colleagues, Maria and Mariana, for their friendship and support during the past five years. I will always remember all the fun and challenging moments we went through together. A heartfelt thanks to Manuel for the continuous encouragement and patience throughout this process. Last but not least, my family deserves endless gratitude. Thank you to my father, mother, brothers and sister who endured this long process with me, always offering support.

Resumo

A experiência do doente é um conceito multidimensional que depende da interação do doente com os sistemas de saúde, as suas preferências, educação, acesso aos cuidados de saúde, etc. A avaliação da experiência do doente tornou-se uma medida habitual para avaliar a qualidade dos serviços de saúde. Os questionários estruturados são as formas mais comuns de medir a experiência do doente nestes contextos. No entanto, neste tipo de questionários, perde-se a profundidade e a complexidade da perspetiva do doente, reduzindo-a a uma simples resposta. Por outro lado, a análise de respostas a questionários ou entrevistas de perguntas abertas é uma tarefa demorada e fatigante: em primeiro lugar, os dados qualitativos podem ser altamente detalhados e extensos e em segundo lugar, devido à sua natureza subjetiva.

Recentemente, vários estudos pretenderam automatizar a análise de dados de experiência do doente em texto livre utilizando o Processamento de Linguagem Natural (NLP). Esta abordagem consiste num algoritmo que transforma a linguagem humana natural para que possa ser processada por modelos computacionais. Assim, torna-se possível explorar dados qualitativos de uma forma mais conveniente e rápida, como por exemplo, a descoberta de palavras ou combinações de palavras que aparecem com mais frequência nos comentários. Além do mais, os modelos de aprendizagem automática supervisionados podem ser treinados com dados de texto classificados e posteriormente utilizados para classificar textos em tópicos ou sentimentos. Por outro lado, os modelos não supervisionados preveem os tópicos ou sentimentos num conjunto de dados sem qualquer conhecimento prévio dos mesmos.

Este estudo faz parte do projeto *CardioFollow.AI*, um serviço de saúde digital financiado pela FCT (Fundação para a Ciência e Tecnologia) e desenvolvido em colaboração entre Fraunhofer-AICOS, NOVA Medical School, Centro Hospitalar e Universitário de Lisboa Central – Hospital de Santa Marta (HSM-CHULC) e VOH.CoLAB. O projeto teve início em janeiro de 2021 e tem a duração de 3 anos. O principal objetivo desta tese é desenvolver ferramentas que captem a experiência do doente em serviços de saúde digital. As duas abordagens utilizadas foram: em primeiro lugar, um questionário de dados estruturados que foi desenvolvido e em segundo lugar, uma técnica de NLP implementada para analisar dados não estruturados da experiência do doente.

Atualmente, não existem questionários validados em português europeu. Portanto, realizámos uma revisão da literatura para selecionar questionários validados em inglês com este mesmo propósito e compilámos perguntas dos questionários escolhidos. As perguntas foram traduzidas e o resultado contém 13 perguntas. O questionário já foi aplicado a 10 pacientes, no entanto, a sua validação terá de ser feita através de um teste piloto de larga escala, de modo a entender se o questionário se adequa à população e finalidade a que se destina.

O principal objetivo do algoritmo de NLP é indicar automaticamente palavras e expressões de um texto que descrevem sentimentos positivos e negativos e identificar os principais tópicos mencionados, permitindo que os investigadores explorem aspetos de melhoria da qualidade. O conjunto de dados utilizados neste estudo consiste em transcrições de vinte vídeos de entrevistas a doentes. Em primeiro lugar, efetuou-se o pré-processamento dos dados: removeu-se a pontuação, palavras irrelevantes e substituiu-se cada palavra pela sua forma básica. A frequência de palavras foi calculada para fins de análise exploratória dos dados. Os termos mais frequentes não permitiram tirar nenhuma conclusão acerca dos temas do documento. Por esse motivo, utilizou-se a técnica designada Part-Of-Speech tagging (POS tagging) de modo a identificar apenas os adjetivos e verbos mais comuns, sendo que os adjetivos mostraram ser os mais úteis para uma análise inicial da experiência do

doente. Posteriormente, os dados foram vetorizados com a frequência do termo-inverso da frequência nos documentos (TFIDF) para captar as palavras mais relevantes no conjunto de dados e para cada doente. Os resultados obtidos permitiram obter mais perceções acerca do conteúdo dos dados e sugeriram que o feedback do doente fosse maioritariamente positivo.

Duas das abordagens foram implementadas para obter o sentimento associado a um conjunto de dados global e às respostas a cada pergunta. Em primeiro lugar, foi desenvolvida uma calculadora que compara as palavras de uma porção de texto com um dicionário que contém palavras associadas a sentimentos positivos e negativos. Consequentemente, indica se o texto inclui mais ou menos palavras positivas ou negativas, indicando assim o seu sentimento. A precisão obtida, quando comparada com dados classificados manualmente, foi de 68%. A baixa precisão pode ser devida à inadequação do dicionário fornecido à calculadora e porque este método não considera a semântica de uma frase, mas apenas significados de palavras individuais.

Em segundo lugar, dois algoritmos de aprendizagem automática, *Support Vector Machine* (SVM) e Naïve Bayes (NB) foram treinados e testados com o conjunto de dados previamente classificado manualmente. Posto isto, como o conjunto de dados estava altamente desequilibrado devido ao baixo número de exemplos negativos, aplicou-se uma técnica de sobre amostragem de modo a equilibrar as classes. O TFIDF foi utilizado novamente para vetorizar os dados, e a exatidão final foi de 59% para NB e 78% para SVM.

Por fim, um algoritmo não supervisionado, *Latent Dirichlet Allocation* (LDA), foi criado para prever os tópicos desse conjunto de dados. Esta abordagem é extremamente benéfica porque não limita as informações a tópicos pré-determinados e permite que o modelo descubra tópicos não previstos pelo investigador. O sentimento de cada tópico foi calculado, usando a abordagem baseada em dicionário. Ainda que as palavras que constituíram os tópicos finais não tenham sido muito coerentes entre si, foi possível inferir o tema de alguns tópicos. O sentimento obtido para a maioria dos tópicos foi positivo. Os únicos aspetos associados ao sentimento negativo, são relacionados com a recuperação física da cirurgia e não relacionados com o serviço digital.

A amostra de dados era muito pequena para esperar bons resultados da análise de NLP. Além disso, a qualidade do conjunto de dados também não era ideal. Assim, um conjunto de dados mais extenso de doentes seria necessário para avaliar a viabilidade do modelo. Por outro lado, existem várias ferramentas que oferecem tarefas básicas de NLP para a língua inglesa, no entanto, existem poucas soluções referentes à língua portuguesa, sendo que as existentes apresentam desempenho inferior quando comparadas às inglesas.

De qualquer modo, aplicando ferramentas de análise de texto, foi possível demonstrar como uma metodologia de NLP pode ser utilizada numa escala mais elevada para dados em português europeu. Em primeiro lugar, o NLP facilita a análise de texto ao descobrir as palavras mais relevantes em documentos através de estatísticas TFIDF, que mostrou ser uma métrica melhor do que apenas o cálculo da frequência de palavras. A abordagem baseada em dicionário teve um resultado razoável, no entanto, a criação de um dicionário mais completo e adequado aumentaria sua precisão. Os modelos de NB e SVM mostraram ser uma abordagem eficaz para a classificação de documentos de acordo com seu sentimento e de uma forma rápida. Seguidamente, os resultados de precisão obtidos foram pobres, todavia, um conjunto de dados mais extenso melhoraria esse desempenho.

Concluindo, este trabalho contribuiu para o desenvolvimento de ferramentas que analisam dados não estruturados da experiência do doente. Adicionalmente, propomos um questionário que facilita a recolha de dados estruturados em serviços de saúde digitais. Os resultados deste estudo

podem ser utilizados para avaliar a experiência do doente em futuros estudos de saúde digital em português europeu. O principal objetivo para futuros trabalhos é otimizar o processo e torná-lo generalizável para qualquer conjunto de dados qualitativos de experiência do doente. Dessa forma, ao inserir texto de transcrições ou de questionários, o modelo retornaria uma representação visual dos resultados, facilitando assim a análise dos dados por parte de um investigador.

Palavras-chave: Análise qualitativa, Processamento de linguagem natural, Análise da experiência do doente, Aprendizagem automática, Classificação de tópicos.

Abstract

Evaluating patients' experiences has become a standard measure to judge the quality of care. However, studying patient experience data may be a time-consuming task. Efficient analysis techniques to examine such feedback have not been frequently applied in European Portuguese, especially for digital healthcare services.

To fill this gap, two approaches were considered. Firstly, for structured data, we compiled and translated items from validated questionnaires for digital healthcare services in English, resulting in a 13-item questionnaire in European Portuguese.

Secondly, a Natural Language Processing (NLP) pipeline was developed to analyse unstructured data. The pipeline was applied to 20 patient interview transcripts from a digital healthcare service. The data was pre-processed, vectorized, and each word was assigned to its morphosyntactic category. Posteriorly, a dictionary-based approach was implemented to return the sentiment of a comment based on the number of positive, negative, or neutral words. Two machine learning algorithms were trained and tested with a previously labelled dataset to classify comments according to their sentiment. Finally, a topic model unsupervised algorithm, Latent Dirichlet Allocation, was created to predict the topics of the text. The sentiment of each topic was calculated using the dictionary-based approach. The most common adjectives gave valuable insights into the text. The obtained accuracy for the sentiment dictionary-based approach was 68% when compared to manual labelling, and 59% and 78% for each machine learning model.

The results made us conclude that the overall satisfaction with the project was positive. However, a larger dataset would be necessary to examine this model's feasibility, as the study's main limitation is the dataset's size and quality. Still, by applying text analytics tools, we could demonstrate how NLP could be used on a larger scale for European Portuguese datasets to maximize the usefulness of patient feedback data and reduce the time invested in its interpretation.

Keywords: Qualitative Analysis, Natural Language Processing, Patient experience analysis, Machine Learning, Topic Modelling.

Contents

List of Figures

List of Tables

List of Abbreviations

BOW - Bag of Words

CSQ - Client Satisfaction Questionnaire

FCT - Fundação para a Ciência e Tecnologia

HSM-CHULC - Hospital de Santa Marta, Centro Hospitalar e Universitário de Lisboa Central

IoT – *Internet of things*

LDA - Latent Dirichlet Allocation

LSA - Latent Semantic Analysis

ML - Machine Learning

NB - Naïve Bayes

NLP - Natural Language Processing

NMF - Non-negative Matrix Factorization

NMS - Nova Medical School

NPMI - Normalized Pointwise Mutual Information

PACT - Patient Assessment of Communication during Telemedicine

PoS - Part of Speech

PREMs - Patient Reported Experience Measures

PROMs - Patient Reported Outcome Measures

PSQ - Patient Satisfaction Questionnaire

QUIS - Questionnaire for User interaction Satisfaction

RMSS - Remote Monitoring Satisfaction Survey

RPM - Remote Patient Monitoring

SMOTE - Synthetic Minority Oversampling Technique

SUS - System Usability Scale

SUTAQ - Service User Technology Acceptability Questionnaire

SVM - Support Vector Machine

TAM - Technology Acceptance Model

TFIDF - Term-Frequency Inverse Document Frequency

TSUQ - Telemedicine Satisfaction and Usefulness Questionnaires

TUQ - Telemedicine are the Usability Questionnaire

UEQ - User Experience Questionnaire

VOH- Value for Health

WHO - World Health Organization

Chapter 1

1 Introduction

1.1 Motivation

Over the last couple of decades, patient perspective has become widely recognized as one of the leading indicators of the quality of healthcare services. Not only has it become a key strategic focus for healthcare providers, but investigators have also introduced various metrics to measure it[1–6].

Although patient experience can be collected through questionnaires and surveys, semi-structured in-depth interviews are the most frequent qualitative data source in health services research. This method typically involves a dialogue between researcher and patient, guided by a flexible interview protocol. By collecting open-ended data, researchers can explore the participants' thoughts, feelings, and beliefs about a particular topic and delve deeply into personal and sometimes sensitive issues[7].

Analysing interview responses requires substantial effort due to the unstructured nature of the answers. Raw free-text data are, therefore, not constantly analysed systematically, risking the loss of potentially valuable insights for hospital improvement. Thus, there is increasing interest in applying Natural Language Processing (NLP) techniques to generate structured data out of texts from large datasets automatically.

Today's text analytics can analyse more language-based data than humans, without fatigue and in a consistent, unbiased way. NLP is a specialized field of computer science and engineering concerned with designing applications and systems that enable the interaction between machines and natural languages used by humans[8]. NLP can discover 'topics' occurring in a collection of documents (*i.e.*, topic modelling) and also detect the sentiment of a topic, assigning a response with a sentiment score. A combination of both sentiment analysis and topic modelling can make optimal use of patient feedback responses. Various algorithms have been created for this purpose for English-based text analysis, but very few are available for Portuguese[4–6,9–13].

The area of NLP is in deep expansion in many countries, and Portugal is no exception. Today several companies have projects in this field, developing data search systems for doctors, translation systems, etc[14]. However, even though there are some pre-trained tools for the European Portuguese language, most of the available open-source libraries for NLP are for the English language or for Brazilian Portuguese. We see that sometimes there are no immediately available solutions, and the existing tools only perform a fraction of the tasks that English-based algorithms perform. It is also worth noting that, even when these tasks exist for the Portuguese language, the performance is usually lower than for English[15].

This research work was developed in Value for Health CoLAB (VOH.CoLAB), and supported the research activities of the *CardioFollow.AI* project, funded by Fundação para a Ciência e Tecnologia (FCT). I chose this topic for my masters thesis because I wanted to undertake a project applied to a real-life context in digital healthcare. The *CardioFollow.AI* project

particularly caught my attention, and I was eager to contribute to its success. I aimed to address a practical need within the project and make a meaningful impact on digital healthcare services.

1.2 Thesis objectives

The main objective of this thesis was to develop tools to collect and analyse patient experience from a digital healthcare service.

Two approaches were considered:

- The development of a survey-based instrument to capture structure data on patient experience of digital healthcare services.
- The implementation of a NLP pipeline to efficiently capture patient experience from semi-structured interviews in Portuguese language. This was done by developing methods to automatically detect words and expressions that describe positive and negative emotions, and also to identify the main topics that are mentioned, that is, through sentiment analysis and topic modelling.

The research aims to answer the following questions:

- What are the existing survey methods for capturing patient experience with digital healthcare services? How suitable are these methods in the context of healthcare services in Portugal?
- Can qualitative analysis of semi-structured interview responses, namely NLP, be used for analysing patient experience with health services in the Portuguese language?

1.3 Thesis organisation

This thesis is organised into six chapters. Chapter 1 briefly introduces the context and motivation for this thesis, along with the corresponding objectives and structure. Chapter 2 presents the study in which this thesis is inserted, as well as fundamentals of patient experience, qualitative analysis, NLP, and Machine Learning. Chapter 3 provides a literature review on the dimensions of patient experience, methodologies used to evaluate patient experience, and the state-of-the-art in NLP methods used to assess qualitative data of patient experience. Chapter 4 describes the primary methods used to create a patient experience questionnaire and the approaches applied to pre-process and analyse the acquired text data from patient interviews. Chapter 5 presents the results of the unstructured data analysis pipeline and discussion. Chapter 6 concludes this thesis with a summary of this work's achievements and gives suggestions for future work. Finally, the last chapter includes the references used for this thesis.

Chapter 2

2 Background

2.1 *CardioFollow.AI* project

CardioFollow.AI [16] is a project funded by FCT (Fundação para a Ciência e Tecnologia) and developed in collaboration between Fraunhofer-AICOS, NOVA Medical School, Centro Hospitalar e Universitário de Lisboa Central – Hospital de Santa Marta (HSM-CHULC) and VOH.CoLAB. The project started in January 2021 and has the duration of 3 years.

This project enabled the implementation of a remote patient monitoring (RPM) system in HSM-CHULC for cardiac surgery patients, which are highly susceptible to severe postoperative complications. The project started during the pandemic crisis, in which patients identified as a risk group were advised to stay home and avoid hospital visits due to the high risk of infection. With the RPM system, the clinical team could remotely follow-up with many patients at home, reducing hospital visits without affecting the quality-of-care delivery. Furthermore, in the scope of cardiothoracic surgery, the potential of telemonitoring systems to support patients' follow-up remains unexplored.

Using an IoT system, patients remotely collect daily outcomes to complement and improve the current follow-up process, consisting of periodic phone calls and consultations over the first year after the procedure. The IoT system consists of a weight scale, a sphygmomanometer, smartwatch, and mobile phone that patient uses to self-report their symptoms through a chatbot. Through this chatbot, patients can receive advice from clinicians and take a picture of their operation wound through their mobile phone. All devices are connected by Bluetooth to the mobile app called *SmartBeat*.

A Machine Learning (ML)-based decision support module is being developed to support the identification of patients with higher risk of complications in the post-surgery period. The data collected from patients' health pathways will determine risks of complications throughout the follow-up process, namely: estimate in the pre-surgery period, optimal follow-up resources; identifying patients who will benefit the most from telemonitoring; and early detection of complications at home, leading to rapid medical intervention.

The research team carried out a pilot study with 35 patients using the RPM system during 30 days after the surgery. These patients were interviewed when they handed in their IoT kits after recovery and asked to complete a satisfaction questionnaire. Currently, a clinical study is being implemented at HSM-CHULC, enabling continuous monitoring and optimisation of ML modules. It is estimated the involvement of 300 patients. Furthermore, in these digital transformation projects in health, the participation of patients and health professionals is critical to impacting health outcomes. As a result, there is a need to comprehensively evaluate patients' experiences and based on their feedback, enhance the effectiveness of digital health services to better meet the specific needs and preferences of individual patients. Thus, the clinical study aims to evaluate the experience of the patient and healthcare professionals and the importance of this system.

This thesis is part of this project, aiming to investigate methodologies for evaluating the experience of interaction of patients and health professionals with the digital service, and also to generalize it for other digital healthcare projects.

2.2 Digital Healthcare concepts

According to the World Health Organization (WHO), telehealth is the delivery of healthcare services where patients and providers are separated by a distance[17]. Telehealth is usually interchanged with telemedicine, but its important to understand that these terms refer to different ways of using technologies to deliver health care services. Although there are no precise, unique, or definitive definitions, Figure 2.1 shows a commonly accepted framework representing the relationships between these terms established by the Joint Action to support the eHealth Network[18]. eHealth is the broadest definition for all technological services. It involves all activities that use communication and information technologies to store, retrieve, share, and exchange health-related information for prevention, diagnosis, treatment, monitoring, educational, and administrative purposes[19]. Telehealth is a subpart of eHealth that can refer to clinical and non-clinical services, such as provider training, executive meetings, and telenursing.

Telemedicine is a part of telehealth once it only applies to the dynamic, real-time interaction between medical providers and individuals seeking health services. Examples of telemedicine are teleradiology, telepathology, telesurgery, and telecardiology[20]. There are two types of telemedicine services: live interactive videoconferencing (synchronous), the store-and-forward transmission of medical images/information, or remote monitoring (asynchronous). Live telemedicine involves real-time, two-way communication between a patient and provider and is mainly used for ambulatory subspecialty consultations. Asynchronous telemedicine consists of a specialist's review of recorded health history with digital images of the video[21].

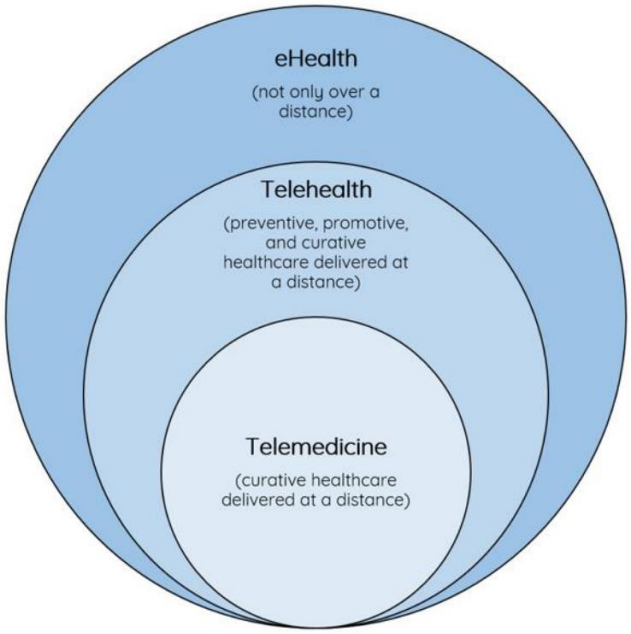


Figure 2.1 - Relationships between the terms eHealth, Telehealth and Telemedicine.[18]

2.3 Patient experience concepts

In patient-centred care, there are two concepts worth mentioning as they are often mistaken: Patient-Reported-Experiences Measures (PREMs) and Patient-Reported-Outcomes Measures (PROMs). The PREMs are tools that capture what happened during an episode of care and how it happened from the patient's perspective[22]**.** On the other hand, the PROMs aim to measure patients' health status and include measures related to symptoms, functional status, and health-related quality of life. These are typically used to monitor patient progress, the effectiveness of treatment or surgical intervention, etc[23].

The term "patient experience" is also mistaken for "patient satisfaction"**.** Patient satisfaction is an outcome measure that combines patient experience of care and health outcomes, indicating whether the care provided has met the patient's needs and expectations. Patient experience is precisely the process indicator that depends on the interaction with health systems, sociodemographic characteristics, previous health care seeking, clinical history, facility characteristics, etc[24]. For example, while a PREM might be a question asking the patient whether he was given discharge information, a patient satisfaction question would be if the patient was satisfied with the discharge information they received. Another term used in this thesis is “usability”, which affects patient experience but refers solely to a user’s interaction with digital interfaces used in healthcare. For example, the digital interface design and how pleasant a system is to use for the consumer is part of the “usability”[25].

2.4 Qualitative analysis concepts

Denzin and Lincoln defined qualitative data as: "a set of interpretative, material practices that make the world visible. These practices transform the world. They turn the world into a series of representations including field notes, interviews, conversations, photographs, recordings, and memos to the self"[26]. Qualitative research is designed to recognize the thoughts and experiences of individuals being studied about a specific issue. There is a more significant focus on the interpretation of a concept and the social phenomena behind it rather than simply accepting the concepts; it aims to answer the questions "what", "how," or "why" about social aspects of health, illness, and healthcare[27].

Natural sciences, and most quantitative research, are typically aligned with a *positivist paradigm* which states that research should be objective, and methods must be unbiased and rational. However, this perspective can be somewhat challenging in social sciences and humanities research. As a result, qualitative research follows an *interpretive paradigm* that argues that research should try to play an unbiased part in understanding the perspective of the world and those who live in it[27].

Qualitative data can be analysed quantitatively. This means that one can apply an unambiguous and predefined coding system to this kind of data, *i.e*., codes such as “positive”, “negative”, and “neutral” sentiment, and posteriorly derive conclusions through frequency measures, *i.e*., counting the number of times a specific keyword in each code[27].

In the process of coding, there are two types of scientific reasoning: ***inductive and deductive methods***. The latter refers to research methodologies that start with a pre-determined theory or hypothesis and proceed to study that theory, confirming it or disapproving it. In contrast, an inductive approach searches for patterns within the observations and states the hypothesis while the study advances. This scientific reasoning can be applied to coding, for example, in inductive

approaches, one adds coding themes as they read the data. Furthermore, it can be applied to interview development questions (in inductive methods, researchers add questions as they are interviewing)[27].

In health research, it is hard to state that a method is entirely inductive, as researchers tend to incorporate their previous knowledge on a specific topic to shape a research question or a coding scheme. For that reason, it is common to use the term "***abductive logic***" to describe a middle position that uses both deductive and inductive methods; for example, starting with some research questions but remaining flexible to include more themes that appear relevant throughout the study[27].

There are several ways to improve validity in qualitative research. For example, ideally, coding should be performed by at least two researchers after defining a common approach. However, this method raises a problem of achieving consistency. It is considered best practice to report ***inter-annotator agreement*** for a corpus. One way of measuring this is using the ***Kappa coefficient***, shown in equation (2.1) which expresses the level of agreement between two annotators in a classification problem. If we ask two raters to classify objects into categories one and two, as shown in Figure 2.2, one could calculate the observed level of agreement, $p_o = p_{11} + p_{22}$. Also, by calculating $p_e = p_1 p_1 + p_2 p_2$, we obtain the expected agreement when both annotators assess the labels randomly and independently.

		Rater 1		
		1	2	Total
Rater 2	1	p_{11}	p_{12}	p_1
	2	p_{21}	p_{22}	p_2
	Total	p_1	p_{12}	1

Figure 2.2 - Example of an object classification by two independent raters. p_{11} is the probability of rater 1 to choose object 1, p_{21} is the probability that rater 2 chooses object 1, etc.

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{2.1}$$

Triangulation is another example of assessing quality of a qualitative study: this method compares results from two or more methods of data collection, for example, interviews and questionnaires. It relies on the assumption that strengthens one method will compensate for any weaknesses in another method. ***Member checking*** **or** ***respondent validation*** is another possible method in which the researcher's interpretation after coding is compared with those investigated.

Reflexivity is a fundamental concept in this topic. It means sensitivity to how the researcher might be shaping the collected data due to prior knowledge, assumptions, experience, and opinions. It is recommended that researchers keep a personal research diary in which they keep their reactions and make their biases plain and available outside of the researcher to enhance the credibility of their findings[28].

2.5 Natural Language Processing (NLP) concepts

NLP is defined as 'any computer-based algorithm that handles, augments, and transforms natural language so it can be represented for computation[5]. It aims to support systems that can

communicate in the same way as humans, that is, using the same language humans use to speak and write. NLP algorithms are usually implemented in a chain or pipeline, meaning that the output of one task will be the input of another task. To understand this pipeline, it is essential to understand some concepts about processing language, as described in the following subsections.

2.5.1 Processing Text

The first step of pre-processing text is text normalization, which means converting text to a more convenient, standard form[29]. Initially, textual data is split into smaller parts called ***tokens***. Tokens are usually made up of words but can be sentences or paragraphs and are commonly used as a starting point for the pipeline. Generally, tokenization can be defined as breaking down textual data into smaller meaningful components.

Another step of normalization is trying to make the text more uniform to facilitate the association of words that, despite having different word forms, have the exact origin and most likely have the same meaning. There are two ways of doing this: ***stemming*** and ***lemmatization***. The latter refers to identifying two words with the same root; for example, the ***lemma, or*** the ***root word,*** of the word "*eating*" and the word "*ate*" is *eat.* On the other hand, stemming is a simpler version of lemmatization, in which the end of the word is simply eliminated, and only the ***root stem*** of the word is considered. So, for example, if we take the word "jumping," the word "*jumped*," and the word "*jumps*," they all have the same root, "*jump*," if we remove its ending[29].

Just like many other intelligent systems, each module of an NLP system, such as lemmatization and stemming, are developed based on a reference dataset. These datasets are called ***Corpora*** (or ***corpus***, in a singular form). They are extensive and structured collections of texts or textual data annotated and tagged with metadata, that is, each word is tagged with its respective stem or lemma. Text corpora are manually annotated by professionals experienced in the language or based on crowdsourcing[15,29].

Furthermore, text normalization includes removing ***stopwords*** and case conversions. Stopwords are words that have minimum or no significance, like common words in English, prepositions, numbers and words that do not contain any relevant information for the study[30]. Examples of stopwords are "a" and "the", which are part of structuring phrases but aren't relevant for textual analysis. Case conversion is simply lower casing all the text.

Once the text is normalized, several types of analysis that can be done. One example is to organize the text by ***n-grams***, in order to have some context in which words appear. These are a sequence of n words: a 2-gram (which we'll call bigram) is a two-word sequence of words like "*please turn*", "*turn your*", or "*your homework*", and a 3-gram (a trigram) is a three-word sequence of words like "*please turn your*", or "*turn your homework*"[29].

Another example, a common approach to understand textual data, is looking at it as grammar: a set of rules that aim to understand how antecedence and provenance of certain words affect others. This approach is called ***Part-of-Speech Tagging*** (*PoS Tagging*). It's the operation responsible for establishing a word's relationship with those that precede and following it to establish a connection with the remaining terms of the sentence. Figure 2.3 shows a sentence, used as input in a PoS tagging function, and its output vector, which contains each word and it's respective grammatic function ('Det' stands for 'determinant', 'Adj' for verb, etc.)

```
sentence = 'The brown fox is quick and he is jumping over the lazy dog'
[('The', u'DET'), ('brown', u'ADJ'), ('fox', u'NOUN'), ('is', u'VERB'),
('quick', u'ADJ'), ('and', u'CONJ'), ('he', u'PRON'), ('is', u'VERB'),
('jumping', u'VERB'), ('over', u'ADP'), ('the', u'DET'), ('lazy', u'ADJ'),
('dog', u'NOUN')]
```

Figure 2.3 - Example of a sentence and its PoS tagged output[31]

2.5.2 Text classification

Classification is the task of choosing the correct class label for a given output. One of the most relevant and demanding problems is text classification or categorization, which involves organizing text documents into categories based on each text document's inherent properties or attributes. This is where techniques like feature extraction and supervised, or unsupervised ML come in handy. Document classification is a generic problem, not limited to text alone, but it also can be extended to other items like music, images, video, and other media[8].

As represented in Figure 2.4, ***supervised models*** involve predetermined output attributes, in which the algorithm attempts to predict the targets. These are the most common models in text classification models. Conversely, ***unsupervised learning*** involves pattern recognition without the involvement of a target attribute. Instead, these algorithms identify associations within unlabelled data and assign labels to each data value[5]. Other relevant concepts in ML are:

- ***Model***: Built using a combination of data/features and a ML algorithm that could be supervised or unsupervised.

- ***Feature extraction***: The process of extracting valuable features, usually in numerical form, from raw data that are used to train ML models.

- ***Features***: Various valuable attributes of the data (examples could be age, weight, and so on for personal data).

- ***Training data***: A set of data points used to train an ML model. In supervised algorithms, the model learns from the features and tries to infer patterns leading to a specific outcome. After the training stage, the model is expected to be generalized to predict classes for new data points in the future.

- ***Testing/validating data***: A set of data points on which a pre-trained model is tested and evaluated to see how well it performs. The test dataset is representative of what a new data sample might be.

- ***Hyperparameter tuning/optimization:*** After evaluating the accuracy scores, hyperparameters can be adjusted to optimize the mathematical function that defines the model. These are called hyperparameters because they cannot be learned from the data but have to be adjusted previously[8].

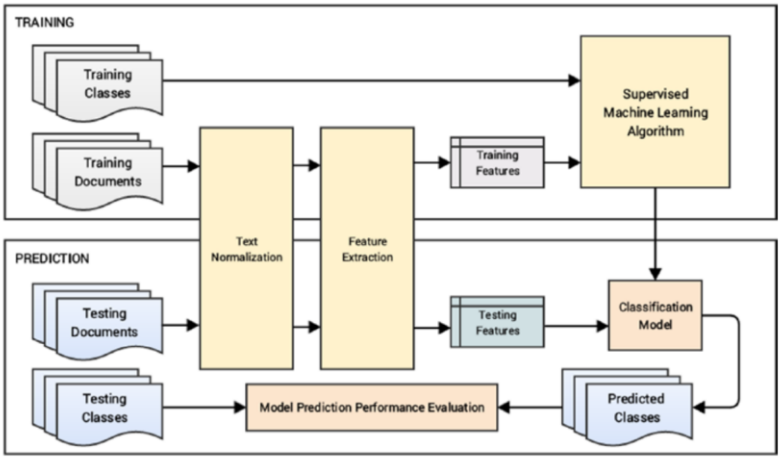


Figure 2.4 - Framework of an automated text classification system[8]

2.5.2.1 Feature Extraction

The big challenge with textual data is that it is highly unstructured and doesn't follow or adhere to structured or regular syntax and patterns—hence we cannot directly use mathematical or statistical models to analyse it. For example, ML models are fed by features that can be numeric or categoric and are unique, measurable attributes or properties for each data point in a dataset. With these features, ML models learn patterns through mathematical operations. Therefore, features should be in the form of numeric vectors. Is then necessary to transform the text into numerical values. This process is called *feature extraction* or *feature engineering*. In this process, a vector space (VS) model is created, which is a document represented by a vector, and the number of columns of a vector is the number of different words (Wn) in that document, as shown in expression 2.2[8].

$$VS = \{W_1, W_2, \dots, W_3\} \tag{2.2}$$

Hence, a document can be represented as the weight (w_{Dn}) of each word in that document, as shown in expression 2.3. The weight is a numeric value that can be calculated in various ways, discussed further.

$$D = \{w_{D1}, w_{D2}, \dots, w_{Dn}\} \tag{2.3}$$

There are two common approaches for feature extraction: the ***Term Frequency Inverse Document Frequency*** (TFIDF) vectorizer and the ***Bag of Words*** (BOW) vectorizer. The most straightforward approach is the Bag of Words model, which simply converts text documents into a vector that represents the frequency of each different word in that same document. The same thing can be done with n-grams[8].

The TFIDF model is based on the statistical measure, shown in formular 2.5, that assesses how relevant a word is in a collection of documents. It is done by multiplying the number of times a word appears in a document (TF - term-frequency), which is equivalent to the BOW vectorizer, and the inverse document frequency of the word through a set of documents (IDF - inverse term frequency), shown in formula 2.4, where N is the total number of documents, and n_i is the number

of documents containing a term t. TF reflects the term frequency of a specific term in a word t_i and a document d_j. [32]

$$IDF = log\left[\frac{N}{n_i}\right] \tag{2.4}$$

$$TFIDF_{i,j} = TF_{i,j} \times log\left[\frac{N}{n_i}\right] \tag{2.5}$$

The final TFIDF metric we will be considering in this thesis is the normalized version of the matrix, meaning that it will be divided by the Euclidean norm, the square root of the sum of the square of each terms' weight. The higher the TFIDF score, the more relevant that word is in a particular set of documents. The ratio given by the total number of documents and the number of documents containing a word allows less common words to have a higher value than prevalent words. The inverse document parcel is essential because it takes care of words that naturally appear in the English language (*i.e.*, "the" or "a"), which are very common in the English language but are not relevant to analyse topics in a document, for example. For example, an IDF value equal to 1 means that a term appears in every document in a collection. The TFIDF value would be equal to zero and the term should be removed as a candidate for "relevance"[8].

2.5.2.2 Supervised classification

In this subsection, two popular supervised ML models will be explored in detail: the ***multinomial Naïve Bayes mode (NB)***, and the ***Support Vector Machine (SVM) model***.

Multinomial NB is explicitly used for classification tasks where we have more than two classes. This model makes a "naïve" assumption that each feature is independent of the others. NB starts by calculating the prior probability of each label based on how frequently each label occurs in the training set, as shown in Figure 2.5. Then, every feature contributes to the likelihood estimate for each label by multiplying it by the probability that an input value will have that feature. The resulting likelihood score can be thought of as an estimate of the probability that a randomly selected value from a training set would have a set of features and a given label, assuming these are all independent.[31]

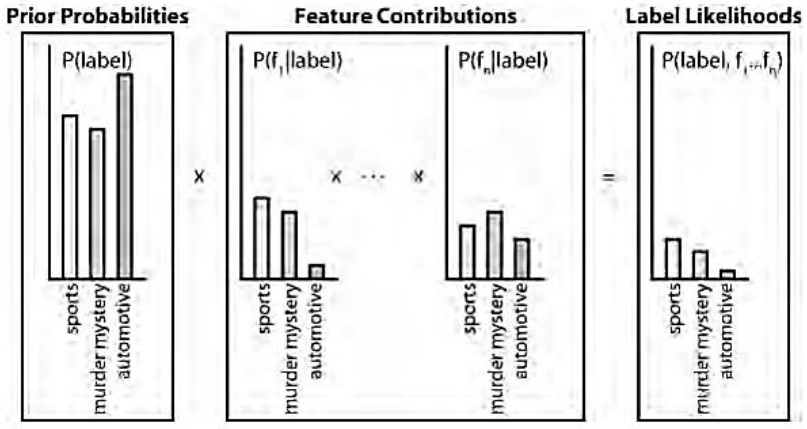


Figure 2.5 - Visual representation of probability calculation with the NB classification model. The box on the left represents the probability of each label, in this case "sports", "murder mystery" and "automotive". The second box represents the probability of the input values with a certain label having a specific feature. The last box is the resulting likelihood, by multiplying the first too boxes[31].

Mathematically, given a response class variable y and a set of n features in the form of a feature vector, using this theorem, the probability of occurrence of a y is[8]:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \times P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)} \tag{2.6}$$

Equation (2.6) can be alternatively written as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \times \prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \dots, x_n)} \tag{2.6}$$

As $P(x_1, x_2, \dots, x_n)$ equally scales $P(y|x_1, x_2, \dots, x_n)$, the denominator can simply be omitted:

$$P(y|x_1, x_2, \dots, x_n) = P(y) \times \prod_{i=1}^{n} P(x_i|y) \tag{2.7}$$

From this equation, a NB classifier can be built by combining it with a *maximum a posteriori* decision rule, where $\hat{y} = C_k$ is a predicted class label, shown in formula 2.8:

$$\hat{y} = \underset{k \in \{1,2,(\dots),K\}}{\operatorname{argmax}} P(C_k) \times \prod_{i=1}^{n} P(x_i|Ck) \tag{2.8}$$

In multinomial NB, the feature vectors are assumed to be word counts from the BOW model, but TFIDF weights will also work. Therefore, knowing that the total number of features in n, and each class label is y, the distribution can be represented as $p_y = \{p_{y1}, p_{y2}, \dots, p_{yn}\}$.

From equation (2.8), $P(x_i|y)$ represents the probability of a feature having an outcome class y. This parameter p can be estimated with a smoothened version of maximum likelihood estimation, described with relative frequency of occurrences, represented in formula 2.9:

$$\widehat{p_{yi}} = \frac{F_{yi} + \alpha}{F_y + \alpha n} \tag{2.9}$$

Where $F_{yi} = \sum_{x \in TD} x_i$ is the frequency of occurrence for feature i in a sample for class label y, and $F_y = \sum_{i=1}^{|TD|} F_{yi}$ is the total frequency for the class label y, in a training set TD.

A SVM model describes the training data sample points as points in space such that either class can be divided by a wide gap between the called *hyperplane*.

The main objective of this algorithm is to find the maximized margin hyperplane that separates the set of data points. In other words, a margin such that the distance between the hyperplane and sample datapoints from each class nearest to it is maximized. Representing this mathematically, considering a training dataset of n datapoints $(x_1, y_1), \dots (x_n, y_n)$ such that the class variable is $y_i \in \{-1,1\}$ where each value indicated the class corresponding to the point $\overrightarrow{x_i}$, a feature vector. Each datapoint $\overrightarrow{x_i}$ is called support vector. Figure 2.6 shows the vector space and hyperplane,

formally defined by $\vec{x}.\vec{w} + b = 1$, or -1, where $\vec{w_i}$ is the normal vector to the hyperplane, and $\frac{b}{\|\vec{w}\|}$ is the offset of the hyperplane from the origin towards the support vectors highlighted in the figure.

When the data is linearly separable, in the current case, we can have margins that are defined by the two parallel hyperplanes depicted by the dotted lines, which help in splitting the data points belonging to the two different classes. This is done considering that the distance between them is as large as possible. The region bordered by these two hyperplanes forms the margin, with the max-margin hyperplane in the middle. These hyperplanes are shown in the figure having the equations.

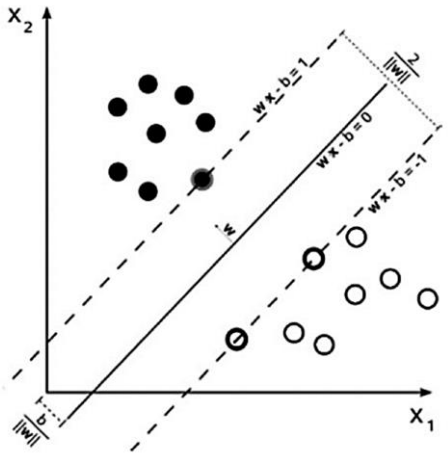


Figure 2.6 - Hyperplane of the SVM model for two classes[8]

2.5.2.3 Evaluating classification models

The performance of classification models is usually founded on how well they predict outcomes for new data points. A ***confusion matrix*** is an ideal measure of performance across the two classes. It consists of a table that helps visualize the performance of classifiers. Each column in the matrix represents classified instances based on predictions, and each row of the matrix represents classified instances based on the actual class labels. For example, we usually keep a class label defined as the positive class, which could typically be the class of our interest. Figure 2.7 shows an example of an accuracy matrix for a two-class classification problem. True Positive (TP) indicates the number of correct predictions for our positive class. False Negative (FN) indicates the number of instances we missed for that class by mis predicting it as the negative class. False Positive (FP) is the number of cases we wrongly predicted as the positive class when it was not. True Negative (TN) is the number of instances we correctly predicted as the negative class.

Accuracy is defined as the overall accuracy or proportion of correct predictions of the model, which can be illustrated by the formula:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.10}$$

Precision is defined as the number of predictions made that are actually correct or relevant out of all the predictions based on the positive class. This is also known as positive predictive value and can be depicted by the formula:

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

Recall is defined as the number of instances of the positive class that were correctly predicted. This is also known as hit rate, coverage, or sensitivity, described by the formula:

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

Finally, the ***F1 score*** is computed by multiplying the previously mentioned metrics as follows:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.13}$$

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Figure 2.7 - Confusion matrix for a two-class classification problem.[8]

2.5.2.4 Imbalanced datasets

Situations occur when the number of instances in one class is much lower than the number of instances in the other classes; this is called the imbalanced dataset problem. To improve classification performance, one possible solution is data-level pre-processing, which operates on the training dataset and changes class distributions using resampling techniques. The most popular approach is oversampling, that is, adding more of the minority class. The inverse strategy involves removing some of the majority class, called under-sampling.[33]

A commonly used approach is a sampling-based algorithm called SMOTE (Synthetic Minority Oversampling Technique), proposed by Chawla *et al.*[34]. It is one of the most adopted approaches due to its simplicity and effectiveness. The minority class is over-sampled by creating synthetic examples rather than oversampling with replacement. The oversampling takes each minority class sample and introduces synthetic examples along the line segments joining all minority class nearest neighbours. The number of neighbours is randomly chosen depending on the amount of over-sampling. For instance, if the amount of over-sampling needed is 200%, only two neighbours from the five nearest neighbours are chosen, and one sample is generated in each direction. Synthetic samples are generated by taking the difference between the feature vector, shown as X in Figure 2.8, and its nearest neighbour, for example, point X2 in Figure 2.8. When this difference is multiplied by a random number between zero and one, it causes the selection of a random point among the line between two features, represented as Y1 in Figure 2.8.

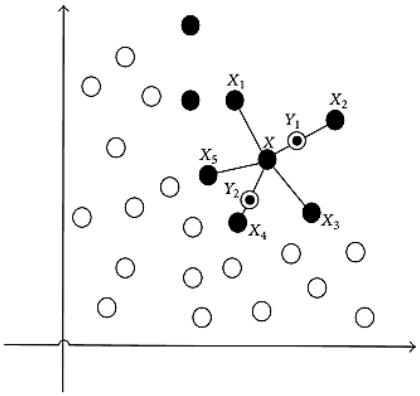


Figure 2.8 - Visual representation of SMOTE oversampling, where y1 and y2 are points generated by the SMOTE algorithm[35]

2.5.2.5 Unsupervised learning

Unsupervised ML algorithms try to discover hidden patterns in data from their features. As mentioned earlier, the data on which these algorithms operate is essentially unlabelled data with no pre-determined category or class. We apply these algorithms to find patterns and distinguishing features that might help us in grouping various data points into groups or topics. The unsupervised technique for text classification, called topic modelling, identifies which topic is discussed in a document or piece of text to categorize patients pre-processed responses into topics without knowing the issues in advance[9]. Topic models can extract critical features from text documents without explicit, manually set rules for information extraction.

In this thesis, the Latent Dirichlet allocation (LDA) model will be the unsupervised ML model used. The LDA technique is a probabilistic model where each document is assumed to be a combination of topics. This method can assemble a set of topics that describe the entire corpus, are individually understandable, and can also handle a large-scale document word corpus[30]. LDA is a BOW model, so it doesn't consider any syntax rules or the order in which the words are written. LDA determines which words are strongly related to each other by counting the frequency with which words appear together in the one document. Other parameters, namely the topics assigned to each word, are hidden or latent.

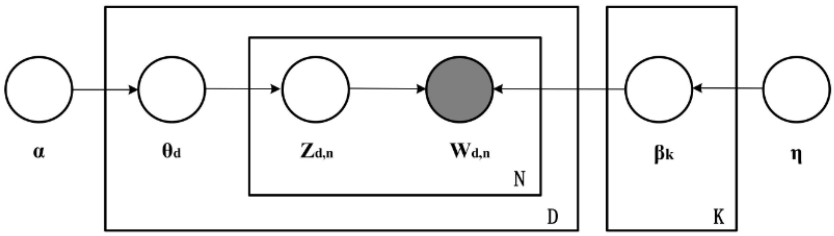


Figure 2.9 – Latent Dirichlet Allocation model representation[36]

In Figure 2.9, the letter W represents each word and Z represents each topic. $\theta(i)$ is then the distribution of topics in a document i. The model also includes two hyperparameters, normally called α and β. α controls the per document topic distribution and β controls per topic word distribution. A high α value means that every document is likely to contain a mixture of most of the topics, and not just a single topic specifically. Similarly, a high β value means that each topic is more likely to contain a mixture of most of the words instead a specific word or group of words. In other words, a

high α value means that documents will be more similar to each other, and a high β means that topics will be more similar to each other, as expressed in equation 2.14. The output of the model is a file with all the topics and words for each topic.

Also, it is necessary to input the number of topics one wants the topic to generate. Determining the number of topics is crucial for text analysis, but how determining the optimal number of topics for the LDA topic model has not been well studied. There are few studies about how many words should be selected to represent each topic. In most studies, this is done by calculating topic coherence or perplexity for each topic and then selecting the one with the better results.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}) = \prod_{i=1}^{K} p(\beta) \prod_{d=1}^{D} p(\theta_d)(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{1:D})) \quad (2.14)$$

- β denotes the topic
- θ denotes the probability of the topic
- z denotes the topic of a particular document or word, and w is the word.
- $\beta_{1:K}$ is the set of all topics, where β_K is the distribution of the words of the kth topic.
- θ_d, is the proportion of that topic in the dth document.
- The set of topics contained in the dth document is z_d where $z_{d,n}$, is the topic to which the nth word in the dth document belongs.
- The set of all words in the dth document is denoted w_d where $w_{d,n}$ is the nth word in the dth document.
- $p(\beta)$ denotes a specific topic selected from the set of topics.
- $p(\theta_d)$ denotes the probability of that topic in a specific document.
- $\prod_{n=1}^{N} p(z_{d,n}|\theta_d)$ denotes the probability that the nth word of the document corresponds tothe topic to which it belongs when the topic is determined.
- $p(w_{d,n} \mid \beta_{1:K}, z_{1:D})$ denotes the joint probability of the topic to which the nth word of the document belongs and the word[36].

The coherence score, shown in formula (2.18) is a widely used performance metric to evaluate topic modelling techniques. It consists of a realistic measure to identify the total number of topics in a document. This measure finds the average word similarity score for each pair of words associated with the topic. The topic model with high Coherence Measure value is considered a good topic model. First, the top frequently occurring words in each topic need to be selected to compute a topic coherence score of a topic model. Posteriorly, the Normalized Pointwise Mutual Information (NPMI) scores are calculated for each of the terms chosen in the first step as well as the coherence score for a particular topic by combining all the Normalized Pointwise Mutual Information (NPMI) scores.

NPMI, shown in formula (2.17) is a measure to find how often two words co-occur in each document. It is calculated by dividing the probability that both words occur by the probability of the occurrence of each word separately[37]. $P(w_i)$ (Equation 2.15) represents the probability of a single word being in document d, 'I' and 'j' are the top words present in topic 't'. $\theta_{w,d}$ represents the word that is present in the document.

$$P(w_i) = \sum_d \theta_{w,d} \quad (2.15)$$

$$P(w_i, w_j) = \sum_d \theta_{w,d} * \theta_{w,d} \tag{2.16}$$

$$NPMI(w_i, w_j) = \frac{\ln P(wi) + lnP(wj)}{\ln P\ (w_i\ , w_j)} \tag{2.17}$$

$$Coherence(t) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} NPMI(w_i, w_j) * P(w_i|t) * P(w_j|t)}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(w_i|t) * P(w_j|t)} \tag{2.18}$$

Chapter 3

3 State of the art

This chapter is focused on providing a literature review of the existing approaches to analysing patient experience across digital healthcare systems. The first section reviews the most common practices to collect and study patient experience in previous studies. The second section briefly describes NLP methods that were developed to process text from patient-feedback interviews or open-ended questionnaires and return insights on patient experience.

3.1 Patient experience in healthcare

Patient-centric care has been defined as "providing care that is respectful of, and responsive to, individual patient's needs and preferences"[38]. Learning more about how patients view the care they receive and how well they address their concerns would be most effective in achieving patient-centred primary care[39]. In the bulletin of the WHO from 2019, it is emphasized that person-centeredness is an essential aspect of quality for two reasons: firstly, it is intrinsically important because individuals have the right to be treated with dignity and respect when they are using health-care services, but also because some studies have stated that patient-centred care is associated with better health outcomes[24].

Doyle *et al.*[40] demonstrated how patient experience is directly linked to clinical safety and effectiveness through a systematic review. The associations between these three aspects were done across various disease areas, population groups, and study outcomes[40]. Therefore, it has become common to judge the quality of care not only by measuring clinical effectiveness and safety but also by gathering patients' perspectives[41]. These metrics can be used as part of systems for developing policy, monitoring the performance of healthcare organizations, and as a mechanism for improving patient experiences and quality care in medical services. Additionally, various patient satisfaction surveys are currently being used to compare clinics, hospitals, and healthcare systems[41,42].

Collected data can be qualitative, quantitative, or mixed (a combination of both). There are also different methodologies for patient experience analysis, such as thematic or statistical analysis. This review encompasses the numerous methods of collecting and analysing patient experience data, as well as the factors contributing to patient experience that vary tremendously throughout the literature.

3.1.1 Dimensions of patient experience

A common concept in service management is "customer journey" defined as the processual and experiential aspects of service processes seen from the customer viewpoint". This concept encompasses the moments before, during, and after a particular service, a continuous measure that makes up a journey. Transcribing this to the healthcare context, the experience, in this case, is not only a matter of interaction with professionals or health outcomes, but a multifaceted experience and many factors may influence the journey of medical treatments[43].

Pickers Institute developed a framework of several dimensions that describe patient-centred care characteristics[44], mentioned and used to design studies by several authors[39,45–47]. Other authors

have proposed their own patient experience surveys and interviews, but these remain consistent with the Picker Institute eight-dimension framework:

- ***Patients' preferences***: providing a respectful environment, always involving patients in decision making, providing the patient with dignity, and respecting the patient's autonomy[46];

- ***Information and education***: giving medical advice[48], providing information about patient's treatment plan[49], alternative treatments[50], and medication comprehensively and understandably, regardless of education level, mother language, age, or health literacy[45,51]. Regarding the moment of informing patients, if the patient is overloaded with information and if they can fully process this information[50]. The frequency of communication with medical practitioners is also relevant to this topic[51], and if there is availability for patients to ask questions comfortably[52]. For example, keeping patients notified and managing their expectations may be an adequate strategy to improve patient satisfaction than decreasing waiting time[6];

- ***Continuity and transition***: if patients feel like there is communication across different healthcare disciplines[45] and if they are provided with explicit information about medications, physical limitations, dietary needs, etc[46];

- ***Family and friends' involvement***: involving family and friends in decision-making, recognizing their needs, for example, providing accommodation if needed[46];

- ***Physical comfort***: management of pain, provision of comfortable facilities[39,45,48]. Hospital environment, that is, cleanliness and quietness are also relevant factors here[46,51];

- ***Emotional support***: if caregivers are attentive to a patient's anxiety about physical status, treatment, and prognosis[46], concern and friendliness shown by a physician[49] and being encouraged to express their worries or sensitive issues[53];

- ***Coordination of care***: posthospital follow-up and support, tracking of clinical tests, communication between health care providers and organizations, and if staff gave conflicting information[39,47]. In the cases of surgical treatments and acute care cases, the way patients perceive the discharge process and how well coordinated it was also commonly considered in the literature as a factor for satisfaction[52,54,55];

- ***Access to care***[48]: ease of making an appointment, the ability of patients to select the day and time of their appointment, availability of transportation[39], availability of treatment needed[50], etc;

Other factors highlighted in the literature are, for example, patient-staff interaction, that is, confidence and trust in doctors and other medical staff[48,54,56], staff responsiveness, and how fast medical practitioners answer patients' needs[51]. Financial aspects are also regularly noted, namely issues related to dealing with medical insurance companies and meeting their requirements or seeking financial advice or consultation[50,57].

3.1.2 Approaches for measuring patient experience data

Creating a data collection tool may be challenging, and one must consider several aspects when planning to measure patient experiences. Firstly, independent of the type of method, researchers must define what needs to be measured; a straightforward question must be aimed at studying well-defined control and dependent variables[58]. Posteriorly, defining the moment when questions are asked is also crucial; for example, interviewing a patient immediately after using the services may be more advantageous as the experience is fresh in people's minds. However, in cases where digital services are used after medical procedures (*i.e.*, surgery) in which patients are unwell, it may be difficult and perhaps unethical to invite someone to share their views instantly after the service ends[48].

Depending on the purpose of a study, it is essential to assess whether it would be helpful to use qualitative or quantitative methods. This section summarizes both approaches, followed by a brief comparison between both.

3.1.2.1 Qualitative measures

The most common qualitative data collection approaches in medical research are interviews. Interviews are used in studies in which understanding individual perspectives of a phenomenon rather than generalizing conclusions from large groups of people[59]. Three types of research interviews are structured, semi-structured, and in-depth. The structured interviewing process is commonly used for quantitative analysis as it consists of closed questions, usually with "yes" and "no" answers, and not much margin for storytelling. Semi-structured interviews are the most common type in patient experience research[45,56,60–67], where there is an interview guide, but it also allows the participant to talk freely about other topics that were not in the initial guide. Furthermore, these interviews should not be conducted with a transactional question-answer approach but should have iterative interactions and conversations between the interviewer and interviewee[68]. Finally, in-depth interviews aim to understand detail about experiences, thoughts, and perceptions; therefore, the interview guide consists of very few open questions[69,70].

Focus groups are group interviews that encourage participants to communicate to generate data. Although focus groups' most apparent advantage is collecting data from several people simultaneously, the most relevant objective is to collect information from group interaction. After a moment of "ice-breaking," participants ask each other questions, exchange thoughts and opinions and comment on each other's experiences and points of view. At the same time, the researcher plays the role of 'moderator'. This method is commonly used in cases where it is not relevant to understand people's knowledge but to examine how they think and why[71]. Typically, focus groups gather 6 to 12 people to discuss specific topics and are audio recorded or videotaped[72]. Examples of discussions are typical “question-answer” sessions or “think aloud” discussions[73].

Observational methods differ from interviews and focus groups as these don’t rely on talking to people. This method consists of a researcher systematically watching people and their social interaction within their everyday activities. It is the primary method of the natural sciences; for example, a biologist watches how cell structures develop throughout time[27]. In participant observations, the observer is part of the observed location, for example, a nurse working in an intensive care unit. In non-participant observations, the observer is “on the outside looking in”, *i.e.*, present in but not part of the situation, trying not to influence the setting by their presence[74].

3.1.2.2 Data Analysis

The data that results from qualitative data collection methods is usually dozens of audio/video tapes of face-to-face interviews. This data must be transcribed to be analysed, and even when using transcription services, the researcher must make sure it is reliable, that is, as similar as possible to the original recording. Transcribing audio and video are not only about what is said. It is also about expressions: sighs, laughs, and lengthy pauses, which may influence the type of analysis that is undertaken. Transcripts are not the only type of data used in qualitative analysis. Researchers also use field diaries or notes made during or after an interview, which help capture immediate impressions, thoughts, and descriptions[75].

As mentioned previously, qualitative data can be analysed quantitatively by coding data and taking conclusion from frequency measures. In general, researchers do not limit themselves to this type of analysis. Qualitative methods are intended to identify subjective topics exhaustively until saturation point is reached, that is, until there are no new additional categories in the data, rather than statistically representing the data[75].

The data must be labelled or coded in a way that captures every possible nuance and not in a way to reduce the data as much as possible. There is software available to ease the coding process such as ATLAS.ti[76], NVivo[77], and MAXQDA[78], that allow users to highlight keywords and auto-code all occurrences. NVivo additionally incorporates ML to accelerate qualitative coding based on annotations. TEXTANNOTATOR provide systems specialized in semantic annotation after learning from human examples.

As mentioned before, there are two ways of developing coding schemes in qualitative analysis; inductive approaches, in which themes emerge gradually as data is being analysed, or deductive, in which the researcher is limited to the coding scheme that was defined initially[75]. Due to the subjective nature of this type of analysis, more than one researcher typically codes the data and then modifies the codes until consensus is reached[45,60,73]. Codes can be either *descriptive*, explaining higher-level concepts, or *in vivo*, where responses are used directly to create codes in studies where patients' vocabulary wants to be conserved[73].

Thematic analysis is the most used approach in qualitative analysis, possibly because it is the most straightforward approach[45,56,60–64,79]. The researcher groups the data into themes and examines all the cases in the study to make sure that each time's manifestations have been considered and compared, and consequently generates patterns from the codes[75]. Usually, authors follow Braun and Clarke's method[80]: familiarizing with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. Researchers that choose this approach tend to follow an inductive approach, in which codes are applied to subsequent interviews and repeated until every coder addresses a final structure[45,56,60,62,64,65].

Another method in qualitative analysis is grounded theory[70]. Is it similar to inductive analysis because the main objective of this theory is to identify categories as they emerge from the data. This process starts with the coding process, just like thematic analysis, but the main difference is that data collection and analysis are iterative. As researchers conclude, they collect more data to test emerging theories. In summary, this method moves backward and forwards between data and theory[75]. Another widely used method is interpretative phenomenological analysis, where researchers seek to understand how people experience and describe their world. This approach explicitly involves two interpretations: researchers add to the participant's interpretation of events;

it is called "double hermeneutic". This approach is not used as much in patient experience; it helps understand personal experiences in a given situation in a small sample[75].

3.1.2.2.1 Quantitative measures

Structured questionnaires are the most common forms of measuring patient experience in medical care. These questionnaires are designed to produce numerical data. Descriptive statistics (mean, standard deviation) and categoric (frequency) data analyses are commonly used for this purpose[42,81,82]. Some articles have used statistical software packages such as the statistical package for social sciences.

Most studies that use quantitative analysis use validated surveys and questionnaires. However, many create new surveys by compiling pre-existing ones, *i.e.*, joining questions from different questionnaires after a literature review to turn them into more appropriate measures for their studies[79,82,83]. Other examinations create questionnaires from scratch without validating them[81,84]. The most common response scale to questions is the Likert scale, a five-point scale of 1-not satisfied and 5-very satisfied. Until recently, researchers mainly administered these surveys on paper; nonetheless, with new technological developments, surveys can be conducted through SMS, messages, e-mail, and other digital resources, allowing researchers to get real-time feedback[85]. Web or e-mailed surveys benefit from reduced administration costs, improved timeliness, and fewer data entry errors[53]. This literature review won't cover the types of quantitative analysis as it goes beyond the scope of the study.

In order to validate a questionnaire, it is necessary to consider numerous aspects. A valid questionnaire must be feasible, that is, the format must be simple and approachable and questions must be straightforward. Also, it must be reliable and precise, in other words, these instruments should not be affected by investigator bias and errors produced by random factors. Furthermore, researchers must assess internal reliability, which reflect the extent to which the questionnaire items that evaluate the same topic are inter-correlated[86].

Another relevant metric is content validity, which assesses the instrument's capacity to measure that for which it is designed, if it contemplates all the aspects relevant to the concepts under study and if it reflects underlying theories or concepts that are intended to be measured. Finally, an instrument can be measured for its capability to measure change in different individuals or through time, called sensitivity to change. Surveys that lack validity will cause interpretability problems when analysing the results[87].

3.1.2.2.2 Quantitative vs. Qualitative

Depending on what kind of study a researcher wants to make, it is essential to assess whether it would be more indicated to use qualitative or quantitative methods. Combining qualitative and quantitative methods might be helpful to balance the specificity and detail of answers. The apparent advantage of using quantitative approaches is that the results are more easily interpretable and faster to study than qualitative data. Also, this type of analysis typically utilizes validated tools to collect data that can be quickly and cheaply used, facilitating comparisons with other studies[41].

However, questionnaires alone may result in the collection of only superficial data, and one loses the depth and complexity of patient experience, which is reduced to a simple answer[41]. Standard survey questions do not allow patients to elaborate on their experiences, often resulting in missed opportunities to understand issues most important to those patients[46]. Qualitative research helps elucidate ideas that may be previously unknown to the researcher or explore a person's values and preferences[24]. In contrast to quantitative methods that draw conclusions from statistical

methods, qualitative analysis also allows researchers to explore matters that may be unique to interviewees and, therefore, gives voice to minorities[59]. For example, in a study evaluating patient experience in a hospital journey[43] when using a survey, the score of preference for room comfort was one of the lowers scores. However, compared to answers given open-ended questions on what could be improved in their experience, patients judged it not to be very comfortable because of the lack of space for movement with orthopaedic aids. Hence, questionnaire items cannot capture nuances that may be highly relevant to analysing patient experience.

Qualitative information's most apparent disadvantage is how detailed and extensive the resulting data is, making coding highly time-consuming compared to quantitative analysis[88], not to mention human resources needed, higher load on the medical staff, and time availability necessary to carry out on-site interviews, for example. Additionally, a significant disadvantage is how subjective and complex data can be for two reasons; firstly, each researcher can interpret the same information differently. Secondly, it is a may be hard to analyse, as satisfaction with care depends on individuals' values, expectations, and experiences. Also, patients' expectations and values are affected by factors related to the health system and outside factors, such as an individual's social identity, and it may be hard to decide which factors to consider[24].

A practical example of this is in English primary care; in types of care that involve higher proportions of younger patients, ethnic minorities, and socio-economically deprived areas, patients reported more negative experiences. The authors studying this population stated that these results could, in fact, mean that they receive worse care, but also because their expectations were different or because they interpreted the questions in survey instruments in different ways[85]. Some methods to overcome this subjectivity include reframing questions to ask the patient to tell them directly what happened instead of rating the experience[24]. A common mistake in this topic is to ask leading questions; for example, instead of asking, "Did you like this experience?", asking, "You liked this experience, didn't you?". The second question already presupposes an answer and will probably induce the interviewee to answer "yes."

There has been an extensive debate among qualitative researchers over whether qualitative and quantitative research can be compared under the same quality criteria. Some have argued that it is not possible because both represent different paradigms and generate different types of knowledge. This is called the *anti-realis*t position, which supports that using mixed approaches in a single study would be inappropriate. On the other hand, the *subtle realist position* defends that all types of research involve researchers' subjective perceptions, so different methods will produce different pictures of the issue being studied. Consequently, there are ways to assess the different perspectives against quality criteria common to qualitative and quantitative: validity and relevance. Assessing relevance may not be indicated in qualitative research depending on the study's goal because this approach generally focuses on unique situations instead of generalizing to broad populations. However, it can be measured against its validity, defined as the extent to which the account accurately represented the social phenomena to which it referred. In other words, this position defends that the means of assessing the quality of research differ due to the research's purpose and not the approach used.

3.2 Patient experience in digital healthcare services

The Bryl institute described patient experience as "the sum of all interactions, affected by a patient's behavioural determinants, framed by digital technologies and shaped by an organizational culture that influences patient perceptions across the continuum of care channelling digital

health."[89]. Undoubtedly, factors influencing patient satisfaction with digital services differ considerably from those used to describe the patient experience for in-person consults. For example, as mentioned before, factors such as environment cleanliness and quietness are no longer relevant in this context. Therefore, it is vital to understand the experiences of individuals using digital health services and how the design of new technologies can affect patient experience.

Digital healthcare services are one of the most prominent examples of change associated with caretaker vs. patient relations in the COVID-19 pandemic. According to investment figures, it is estimated that there are 626 digital health companies active across Europe today, and 63% of them were founded in the past five years[90].

From a patient's perspective, regular in-person doctor visits became risky, and telehealth can mobilize all aspects of healthcare to decrease disease transmission, conduct people to the right level of health care, and ensure safety for everyone. From a medical professional's perspective, freeing up medical staff and equipment required for those who became seriously ill from COVID-19 and avoiding being infected by patients was possible[91]. Given the growing prevalence of telehealth, numerous studies were done to assess patient satisfaction and usability of telehealth in several areas of medicine. However, in a systematic review by Knapp *et al.*[92] to understand the influence of PROMs and PREMs in telemedicine evaluation studies, the authors found that PREMS were used in less than half of these studies and only played a role in the initial phases of the application. In conclusion, the authors stated that telemedicine specific PREMs should be used more frequently to evaluate telemedicine services, both during and after implementation.

3.2.1 Dimensions of patient's digital healthcare experience

Literature suggests that, generally, patients are satisfied with their digital healthcare experience[49,60–63,79,81,84,93–98]. Studies evaluating patient satisfaction with virtual care show that the most mentioned advantages are logistic convenience; accessibility, convenience for patients with increased childcare responsibilities[66], freeing caregivers from accompanying them, time savings, reducing the unnecessary, not having to miss work or travel, reduced costs and consequently reduced stress[49,56,60–63,81,84,95,96]. Of course, telemedicine is much more relevant where access to care is a barrier, for example, for patients that live in rural areas or far from medical establishments. In a study by Sathiyaraj *et al.*[98] to evaluate cancer patient satisfaction with telemedicine during the pandemic, no one rated virtual visits as better than in-person visits, which contradicts previous literature results. The authors state, "a possible reason for this discrepancy may be that this study was conducted in a setting where access to care was not a barrier."

Furthermore, it is to be noted that most recent studies on this patient experience in digital healthcare were carried out during the COVID-19 pandemic or lockdown[61,63,66,67,84,95–98]. Within the pandemic context, people acknowledged the benefits from keeping them and others safe from infection compared to the need to be seen by doctors in-person[61]. Additionally, it is possible that the high level of satisfaction with telehealth could be due to lower patient expectations of general practice during this time, especially when many believed that health services were less available[61]. Thus, patient satisfaction with virtual care under these conditions doesn't necessarily mean that patients prefer virtual care in a scenario where infection exposure isn't an issue. In a study regarding patients' satisfaction with virtual prenatal care during the pandemic, the results showed that even though patients found a general satisfaction with their virtual experience, the majority of women indicated a preference for in-person care under non-pandemic conditions[97].

There are several validated questionnaires for quantitative analysis of patient satisfaction in digital health. According to a study by Hajesmaeel-Gohari[99] *et al.* the most frequently used questionnaires in telemedicine are the usability questionnaire[100] (TUQ), and telemedicine satisfaction questionnaire (TSQ)[82]the Service User Technology Acceptability Questionnaire (SUTAQ)[101]. After these, the Client Satisfaction Questionnaire (CSQ)[102] is also used; however, it is not directly related to healthcare but can be applied to any service. Also, the Questionnaire for User interaction Satisfaction (QUIS)[103], the system usability scale (SUS)[104], and the Technology Acceptance Model (TAM) surveys are more related to the technical aspects of the equipment rather than the general experience with the telehealth service. Furthermore, the Patient Assessment of Communication during Telemedicine (PACT)[105] was used in a study to compare patient-provider communication in a telemedicine context and face-to-face visits. Other than these, there are the telemedicine satisfaction and usefulness questionnaires (TSUQ)[106], the Post Study system usability Questionnaire (PSSUQ). Finally, the patient satisfaction questionnaire (PSQ)[107] only considers the patients experience with medical care, not specifically to virtual care.

Even though there are several validated questionnaires for telemedicine program evaluation, Finkelstein *et al.*[108] identified the need to develop a more specific questionnaire to evaluate their telemonitoring service. Therefore, the authors proceeded to create a survey to measure patient satisfaction of lung transplant recipients with home monitoring, the Remote Monitoring Satisfaction Survey (RMSS)[108] by compiling items from the literature and adapting to their situation, following the guidelines proposed by Demiris *et al.*[109]

The following categories can summarize the most common factors taken into consideration by researchers and in these questionnaires to analyse patient experience in recent digital healthcare research. These dimensions are represented in Figure 3.1:

- ***Service quality:*** support and availability[84,94], ease of getting help in case of technical difficulties[62,64,65], and if the provider has the same availability to listen, answer any doubts, and spend time with the patient compared to normal appointments[61,63]. Another important factor in this topic is the clearness of instructions when explaining the technology[94,110]: if instructions of use are complex or require an extreme amount of mental effort.

- ***Interaction with medical provider:*** impact of technology on communication and interaction quality with the practitioner[67,93,105,106,111]. Patients claim that video or telephone visits have a loss of personalized feel limited by technology and provoke a certain level of awkwardness in conversations[49,60,61,84]. For example, patients diagnosed with cancer, mental health issues, or other serious illnesses may prefer more physical encounters to provide consolation, which can only be done in person[63,98]. In-person appointments are also more appropriate when patients or clinicians must discuss severe or personal issues when receiving/delivering difficult or bad news[56]. Also, a common advantage cited in some articles is that telehealth can increase communication between medical providers and family members[70].

- ***Quality of care:*** perception of quality of care compared to in-person visits[60,81,94,98,100,106], that is, if the quality of care is comparable or superior to traditional hospital visits. The need for hospital visits also depends on disease severity; telehealth works especially

well for primary non-urgent issues. However, patients prefer to consult a doctor in person for more acute or complex issues[61].

A common disadvantage mentioned as a primary limitation of telemedicine is a lack of physical examination, especially in medical specialties where it is more relevant, for example, in physical rehabilitation[84,95,96]. Consequently, telephone and video consult worried patients about not being adequately observed and examined, and many patients would instead be examined appropriately regardless of logistic convenience benefits[61].

F

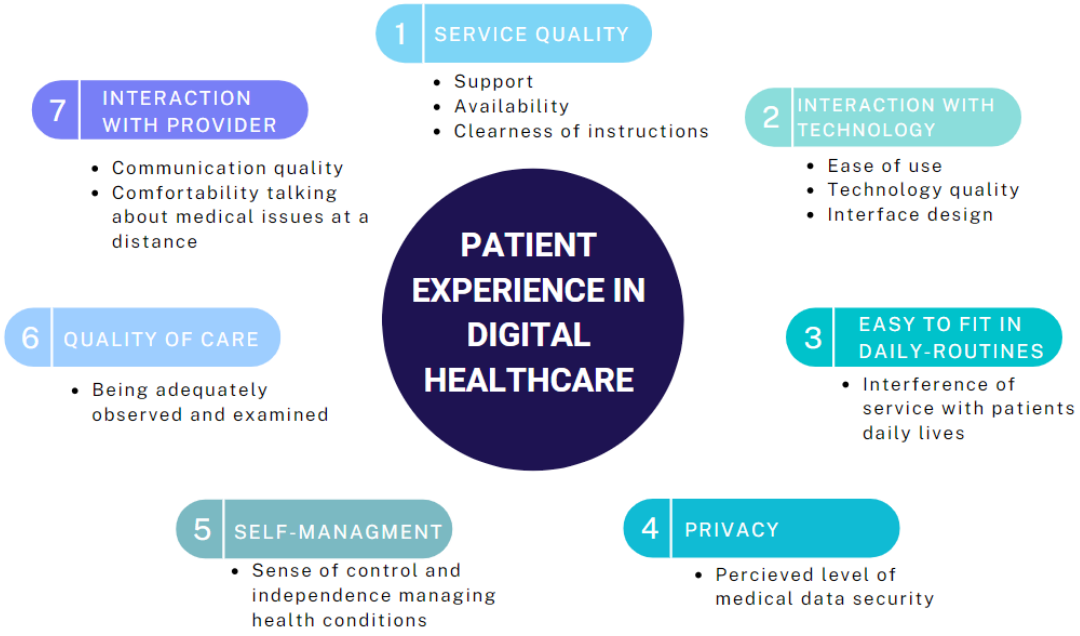


Figure 3.1 - Dimensions of patient experience in digital healthcare

- ***Self-management:*** Sense of control and independence [63,65,94,106,110]; participants expressed increased autonomy by using telemonitoring services and empowered self-management as patients learn and understand more about their conditions. Consequently, patients feel more in control of their health and less concerned. Also, timely interventions and regular monitoring give the patient a sense of reassurance and "peace of mind" as patients know they are being continuously monitored[70].

- ***Privacy:*** level of perceived security and privacy[94,101,106]: if patients have concerns about sending their information across an online platform, or even if patients are uncomfortable talking about their medical issues in other places other than a consult room, for example[61].

- ***Easy to fit into daily routines***[94,110]: if the service interferes significantly with the patient's lifestyle. For example, if measurements made with telemonitoring devices took an unreasonable amount of time or if devices were not transportable, requiring the patient to stay in one location.

- ***Usability:*** the most common disadvantage cited by patients are technical difficulties (*i.e.* video and audio lag, blurry images, and internet connection problems), and ease of

use is also an important factor highly valued by patients[60–62,64,67,70,73,84,93–96,98,100,103,104,106,110,112,113]. In a study evaluating the main decision-making factors when using telehealth at home, the researcher identified the lack of personal competency or negative attitudes towards technology as one of the main barriers[64,66]. The interaction with technology also depends on interface quality, that is, quality of the graphical user interface, ease of navigation and how pleasant the system was to use for the consumer. Furthermore, the system must be reliable, which refers to how easily the user can recover from an error and how the system provides guidance to get back on-track[100]. Another factor is devices comfort and design[110], for example, in the case of wearable telemonitoring devices, some patients valued the appearance of wearing these devices, especially in the case where these are worn in public.

3.3 Natural Language Processing

Recently, several studies have aimed to automate the analysis of free-text patient experience data to inform quality improvements. Qualitative research results in pages and pages of text, making analysis challenging. One could read through patient experience comments or complete video transcriptions to study them. Nevertheless, large health systems have large patient volumes, and thus, collecting and analysing large datasets would require a lot of personnel resources to carry out such a time-consuming task.

NLP is currently the most used 'big data' analytical technique in health care[5]. These algorithms make it far more convenient to explore these types of data; they can enhance the analysis of patient feedback by discovering words or combinations of words appearing most frequently in comments. Furthermore, most studies combine ML algorithms and NLP which can help train models to classify comments into specific categories Consequently, it may provide important information inaccessible from numerical analysis alone[4]. Most of the studies of patient experience using NLP using patient feedback comments available online in social media. For example, Lopez *et al.*[114] gathered online reviews of doctors and classified them according to its sentiment. Other studies use patient review datasets from hospitals[9], patient experience databases such as Press Ganey[115], NHS Choices[12,116], or RateMDs[117], and social media databases[13]. Furthermore, NLP can also be applied to unstructured interviews and survey responses[4].

Manually labelled corpora are the most accurate way of training NLP algorithms, however, due to the high effort that this manual annotation requires, there are not many corpora available. This problem is even worse when it comes to less spoken languages, which is the case of the Portuguese language, namely, the European Portuguese language. Cammel *et al.*[9] analysed patient experience data in Dutch with NLP and started by translating the text to English due to the lack of corpora available in Dutch.

A corpus that stands out for the Portuguese language, namely for the PoS marking tasks, is the *Floresta Sintática* (*Syntactic Forest* in English)[118]. This corpus consists of a set of texts in European Portuguese and Brazilian Portuguese and can be divided into four parts[15]:

- *Floresta virgem:* With 1.6 million words obtained from European and Brazilian Portuguese newspapers.
- *Amzónia:* With 3.8 million words obtained from blogs and non-fiction texts, exclusively in Brazilian Portuguese.
- *Selva:* 300 thousand words divided between different modes - written and spoken.

- *Bosque* - With 190 thousand words, it is the smallest subsection of the *Floresta Sintática* and consists of journalistic text in European and Brazilian Portuguese. This corpus is the only one that linguists have completely manually revised.

The first step in NLP is to pre-process or clean the text and only keep what is relevant for the study. In most studies, pre-processing involves tokenizing, removing stop words, stemming, and lower casing. The pre-processing stage is always crucial in any application of NLP, because the textual components that are obtained (*i.e.*, words, phrases, tokens, etc.) form the building blocks that are fed into further stages of application. If text data is not processed properly, it is most probable that the results will be inaccurate[8]. After pre-processing the text, usually, a simple frequency calculation is done to learn what the most common words are. However, considering only single words may be too general, to avoid losing context, bigram and trigram frequency are also calculated[4,9,119]. This approach may already give some indicators of the most prominent topics in the document. As the large dimensionality of the corpus requires a longer computer processing time, reducing the dimensionality is also expected by removing infrequent words[9].

Several toolkits offer the English language's basic NLP tasks (stemming, linguistic analysis, PoS Tagging). Yet we find that no solutions are immediately available regarding the Portuguese language. Some modules are usually provided that can be used to train models for any language. However, these modules do not exist for all tasks, so there are missing elements. It is also worth noting that, even when these tasks exist for the Portuguese language, the performance is usually lower than for English. For python programming language, the two available toolkits for the Portuguese language are SpaCY[120] and NLTK[121]. NLTK also does this and is the best-known NLP tool for Python, and there are, therefore, many contributions made over time to implement more elements. Both these toolkits have PoS tagger, name entity recognition, lemmatization, stemming, and semantic analysis functions for the Portuguese language.

3.3.1 Sentiment analysis

One of the most common text analytics techniques is sentiment analysis, especially when it comes to analysing patient experience[4]. It is a standard text classification tool that analyses an incoming comment and tells whether the underlying sentiment is positive, negative, or neutral. Some packages already classify comments according to their sentiment, such as pattern.nl (Dutch language)[9], TextBlob[11,13], DICTION[117], and TheyStay[12]. The first straightforward approach is the dictionary lookup approach. The algorithm uses a dictionary with positive and negative examples of words, and the sentence is classified as positive if there are more positive words and vice versa[122]. The challenges in using this approach are related to the complexity of the text since humans can express opinions in different ways, e.g., sarcastic, or ironic content can be misinterpreted[116].

Cammel et al[9] combined sentiment and frequency analysis; that is, for each n-gram, its frequency was plotted against average sentiment. With this information, the authors could conclude that there was a need for improvement on frequently mentioned topics with negative sentiment. Besides dictionary-based approaches, there are supervised ML algorithms for sentiment analysis. In this case, the dataset is manually classified and divided for training and testing datasets, or an external labelled dataset is used to train the algorithm. In the first case, more than one annotator is required to label the data, and then Inter Annotator Agreement is calculated using Cohens Kappa Score as labelling can be a very subjective task[115].

3.3.2 Thematic analysis

Most studies use supervised methods for thematic analysis, which can be called topic classification. Here, the researcher chooses the topics in advance, manually classifies the data according to its subject, and then trains the algorithm according to the classified dataset. The advantage of this method is that the researcher can focus on what is relevant to himself; however, a drawback is how time-consuming manual labelling can be and the fact that the researcher must predict topics that may appear in the data.[115]

In practice, literature's most common topic modelling approaches are, Latent Dirichlet Allocation (LDA), the non-negative matrix factorization (NMF) and Latent Semantic Analysis (LSA). LDA is the most popular algorithm in real life applications, as it doesn't require any previous training data, handles long documents and mixed length document. Its main disadvantage is the fact that it requires a predefined number of topics[30,115,117]. Bahja *et al.*[116] combined topic modelling with sentiment analysis, calculating a sentiment score for each topic model.

The NMF algorithm works better with shorter texts than with large text documents. Its input is a word-document matrix, performing dimension reduction and clustering simultaneously. It aims to obtain topics for short text data by using the factorizing asymmetric term correlation matrix, term-document matrix, and bag of words matrix representation. Its main advantage is the fast processing of large amounts of real-time data, and it can extract meaningful topics without prior information or knowledge of the meaning of the data[30].

LSA is another standard NLP technique in which stated terms with similar meanings are closer in terms of contextual usage. It also analyses large quantities of raw text into words and splits them into meaningful sentences and paragraphs. The mathematical approach, called single value decomposition, reduces the dimensionality of the term-document matrix generated by selecting only the most significant singular values and keeping the first columns of matrices. Cosine similarity is then calculated to understand the similarity of different documents, the similarity of the different words, and the similarity of various terms and documents. The main advantage of this model is it captures synonyms of words and doesn't require a robust statistical background. However, it is hard to determine the number of topics depending on human judgment[30].

Chapter 4

4 Materials and Methods

This chapter describes the methods applied in this thesis project in detail. The first section describes the process of developing a semi structured questionnaire to evaluate the patient experience with the *CordioFollow.AI* project and generalize it to other digital healthcare services. Section 4.2 explains the qualitative analysis process, in this case using NLP and ML algorithms, to conclude valuable insights about the patient's experience in the Portuguese language.

4.1 Collecting structured data to evaluate patient's experience with digital healthcare services

When searching for tools to evaluate patient experience with the *CardioFollow.AI* project, the research team acknowledged that the available validated questionnaires in European Portuguese didn't cover the dimensions the team was interested in analysing. The "Valid Scales Directory for European Portuguese 2020"[123] was used to possibly identify the validated questionnaires in Portuguese to implement in the study. Unfortunately, none of them were related to digital healthcare services.

In the first pilot study the team carried out interviews with twenty patients using a validated survey, the User Experience Questionnaire (UEQ)[124] that measures the "user experience of interactive products". However, after the study ended, the researchers identified several problems with this tool; on one hand, it isn't specific to digital healthcare services. On the other hand, the team realized that patients had difficulty understanding the language used in some items and had numerous doubts when interpreting the questions. Therefore, we identified the need to develop a more appropriate questionnaire for this type of study and population.

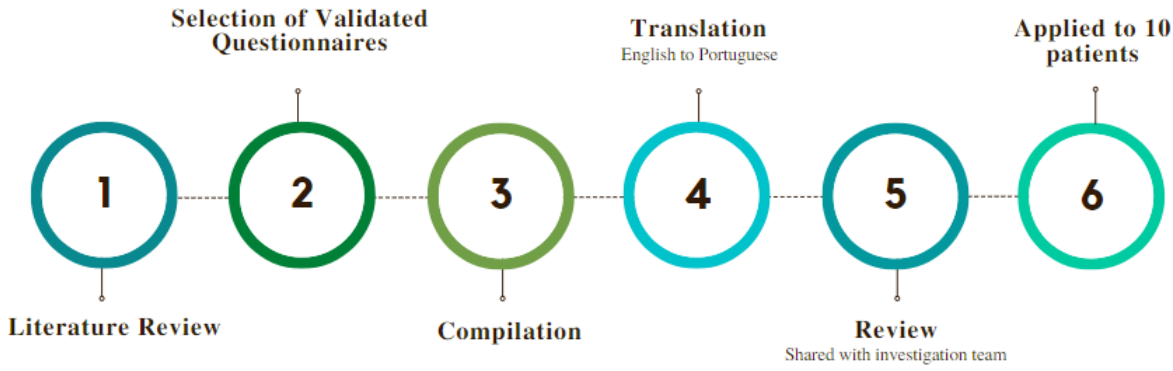


Figure 4.1 - Steps to develop the "Patient experience with digital health services" questionnaire

The process of developing the questionnaire consisted of six phases. The first phase was to conduct a literature review to understand the multiple dimensions of patient experience with digital

healthcare, described in detail in the second chapter. The survey was then developed by selecting validated questionnaires related to patient experience only and that could be applicable to for any digital service and weren't too focused on user-interface and other technical aspects. To make the questionnaire generalizable, some terms in the statements were modified. For example, the term "spirometer", which was the telehealth device used in the study for the RMSS questionnaire, was replaced by "the equipment" (*i.e.*, the statement "the spirometer is easy to use" was modified to "[*the equipment*] is easy to use").

Posteriorly, we followed to merging items from these chosen validated questionnaires. The questionnaire was then shared with other members of the investigation team, who collaborated to remove some redundant or irrelevant items, modified some vocabulary and translated the items to the Portuguese language.

4.2 Analysis of unstructured verbal data using NLP

4.2.1 Toolkit

- *Pandas*[125]: open-source python package widely used for mathematical operations, statistical analysis, loading and saving data, data analytics, ML, and others.
- *NLTK*[121]: The most used framework used for text analytics for Python. It provides several corpora and other resources like stopword lists, tools for pre-processing, frequency distribution calculators, creates n-grams, and others.
- *Scikit-learn/Sklearn*[126]: A free ML algorithm in Python. It also features various algorithms like NB, random forests, and k-neighbours and calculates performance measures. Also, it contains text vectorizing tools.
- *Gensim*[127]: Library used to implement LDA and other topic modelling algorithms.
- *MatPlotLib*[128]: library for creating graphic visualizations in Python.
- *WorldCloud*[129]: Python's library allows the creation of a word cloud that represents text data in which the size of each word indicates its frequency or importance.
- *PyLDAvis*[130]: Tool to visualize the topic model results to ease interpretation.
- *NumPy*[131]: One of the most used python packages for scientific computing. It provides various tools used in math operations.
- *spaCy*[120]: Open-source library for NLP in python, used for PoS tagging in this study.
- *Imbalanced learn*[132]: Open-source licensed library providing tools to deal with imbalanced classes classification problems.

4.2.2 Dataset

The dataset consists of 20 videos lasting between five minutes and twenty-five minutes. Details about the videos are shown in Table 4.1. The videos are recordings of semi structured interviews given to *CardioFollow.AI* study participants in Portuguese after they returned their equipment. In the case where family members were also present, they also spoke about their experiences. Patients were interviewed by the investigators responsible for the study and were performed inside the hospital in various places (appointment room, waiting room, etc.).

The semi structured survey was organized by the following open-ended questions about the patient's satisfaction:

(Q1) What was most important to you in the recovery period?

(Q2) What was most difficult in the recovery period?

(Q3) What would you recommend to improve the service?

(Q4) Was the information given during product delivery sufficiently enlightening?

The videos were manually transcribed verbatim, as none of the automatic transcription tools available were adequate for this purpose. Given a large number of videos, the fact that the participants are speaking in Portuguese, recorded in a rather noisy background, and with an average length of 15 minutes, after an exhausting search of transcription tools, the only efficient option would be to transcribe each one manually. Some artificial intelligence software was tested, but the transcription results were wildly inaccurate. One would have to watch the entire video anyway to correct persistent transcription mistakes using this option. Furthermore, none of the tools found offer an affordable plan for the number of videos needed for transcription.

Table 4.1 - Details about patient feedback interview videos including patients' gender, number of videos, video duration, and number of participants

	Gender	Feedback videos	Duration HH:MM:SS	Number of interlocutors	Number of present caregivers	Number of present nurses	Number of present researchers
1	M	1	0:07:08	4	1	0	2
2	M	1	0:06:02	3	0	0	2
3	F	2	0:07:03	3	1	0	2
4	M	2	0:01:01	2	0	0	2
5	M	3	0:12:00	3	0	0	2
6	M	1	0:09:25	3	0	0	2
7	F	2	0:07:05	2	1	1	1
8	F	2	0:10:38	3	0	0	2
9	M	1	0:08:37	3	0	0	2
10	F	1	0:19:24	3	0	0	2
11	F	1	0:08:03	4	1	0	2
12	F	1	0:08:52	3	0	0	2
13	M	1	0:22:59	5	1	1	2
14	M	2	0:10:46	4	1	0	2
15	M	3	0:05:38	5	0	1	3
16	F	1	0:26:03	4	0	1	2
17	M	1	0:03:06	4	1	0	2
18	F	1	0:13:58	6	1	2	2
19	M	1	0:15:04	5	1	1	2
20	M	1	0:12:48	7	3	1	2
	60% M		**Total time** : 3h25m	**Average**: 4			

4.2.3 Model Design

4.2.3.1 Pre-processing and Exploratory Analysis

Figure 4.2 illustrates a flowchart with various steps of the proposed model, starting from loading the data, cleaning, pre-processing, topic modelling, and sentiment analysis. Each of these steps is explained in detail in this section.

The transcripts were imported into Python's Jupyter notebook and organized by patient document. Then, the interview questions and subject identification names were removed. Patients' unstructured responses to the open-ended questions were initially tokenized using NLTKs interface *RegexpTokenize*. This function separates punctuation and non-alphabetical characters from words, using as input a list of all possible punctuation characters. Posteriorly, using the *simple_preprocess* function from the *genism* library, all the tokens were lower cased. There was no need to include spelling mistakes correction because the transcriptions were revised multiple times to correct eventual mistakes.

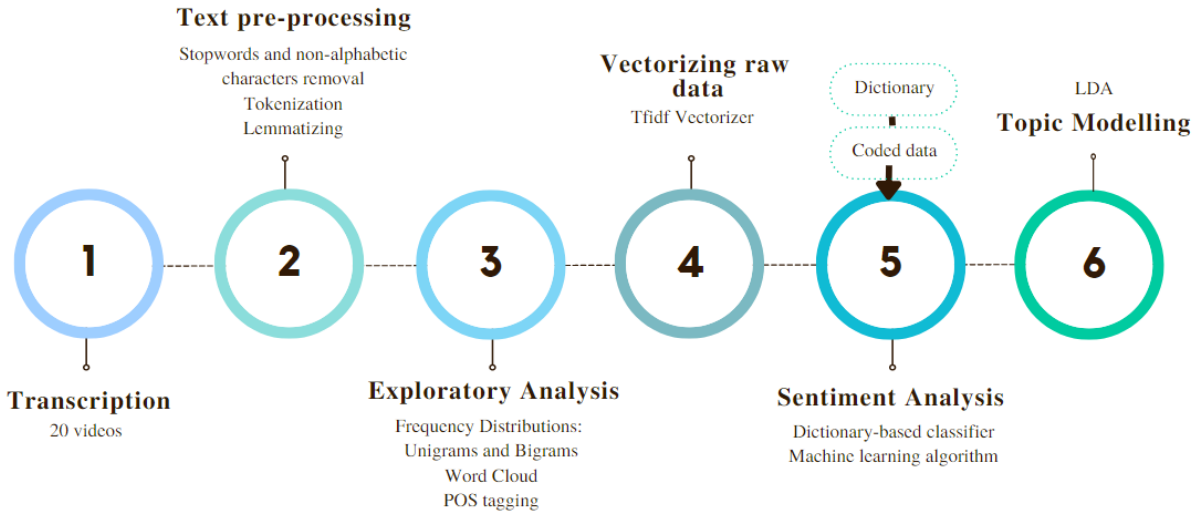


Figure 4.2 - Proposed model framework

Also, the corpus was prepared and cleaned by automatically removing stopwords. Patient names were also removed to keep the answers anonymous. The NLTK package has a corpus for Portuguese stopwords, however this corpus was extended with other words considered irrelevant for the analysis.

Stemming was initially implemented but removed in a later stage as it was affecting the results negatively. We acknowledged that, in many cases, reducing a word to its stem changes its grammatical meaning, as the root stem may not always be a lexicographically correct word; that is, it may not be present in the dictionary. For example, the Portuguese word "*anda*", which means *"walks",* is reduced to "*and*", which has no meaning. Conversely, the lemma, will always be present in the dictionary; the lemma for the previous word is "*andar*", which means walking. For that reason, the data was lemmatized using the *spaCy* library. In this way, we are considering the importance of a word having in consideration every time it is mentioned independently of its variations in verbal tense, for example. This was done by iterating over every token and replacing it with the respective lemma, using the *lemma_* function. In order to use this function, the *pt_core_news_lg* pipeline was downloaded from *spaCy*. This pipeline performs many operations for NLP and is trained with the *Bosque* corpus. Posteriorly, a loop was created to print the words before and after lemmatization in order to assess the lemmatizers' performance.

For exploratory data analysis purposes, the most common words were calculated using the "*FreqDist*" function from the NLTK package and represented in a word cloud. The same thing was done with n-grams, using NLTKs *bigrams* and *trigrams* functions. These functions return n-grams generated by iterating a list of tokens and joining them two by two, for example, if we represent a

list of tokens with numbers, [1,2,3,4,5] , the NLTK bigram function would output [(1,2), (2,3), (3,4), (4,5)]. Similarly, the trigram function would output [(1,2, 3), (2,3,4), (3,4,5)].

Contrary to what it might seem, the most frequent terms did not provide any insight into the topics of the document. For that reason, PoS tokenization using the *spaCy* package was done to identify only the most common adjectives and verbs. This package also uses the *Bosque* corpus to tag the words. Extracting adjectives from answers may tell us how most patients felt about the service Therefore PoS tagging and frequency distributions were combined[4] .

Posteriorly, the corpus was represented with the term frequency-inverse document frequency (TFIDF). The *TfidfVectorizer* function from the *sklearn* package was used for this matter. A list of the tokenized and clean transcripts was used as input. The functions' parameters were specified; the "n-gram range parameter" was set to (1,1) so that the function would output only unigrams and the "*max_df*" parameter was set to 0.80, meaning that the algorithm ignores words that appear in more than 80% of the documents to reduce dimensionality, and also to avoid including irrelevant words. Once the dataset was vectorized, features and scores were extracted from the output matrix and were used to create a data frame and a heat map with the terms with highest TFIDF scores using the *altair* pakcage. The same thing was done with bigrams and trigrams by modifying the "n-gram range parameter" values.

4.2.3.2 Sentiment Analysis

4.2.3.2.1 Dictionary-based classifier

For sentiment analysis, two different approaches were used: a dictionary-based approach and a ML approach. In the first method, a sentiment lexicon containing positive and negative words was necessary. For this purpose, a dataset with positive and negative words in Portuguese was downloaded that was collected by Yanquing Chen and Steven Skiena[133]. After being uploaded into Python, two lists were created, one containing positive tokens and the other negative.

Firstly, the transcripts were organized into a list of patient answers. Each answer was previously tokenized, cleaned, and lemmatized. To calculate the sentiment of an answer, a function that I shall be designating as a "sentiment calculator" was created to iterate over all the patient responses and count the number of positive and negative words in each answer. The function asks for a threshold number for positive and negative terms. If the positive word threshold is set to 1, for example, it means that there must be more than one positive word in the answer for it to be classified as positive. After several rounds of trial and error, we concluded that the best threshold values were 1 for positive and 0 for negative for better results. If an answer was not classified as positive or negative, it is considered "neutral".

Once the whole dataset was classified, to compute the calculator's accuracy, all patient's answers were manually labelled as positive, negative, and neutral by two researchers independently. Data was coded through an open-source text annotation tool called Doccano[134] and uploaded into python. The Kappa Coherence score was used to calculate inter-rater coherence using the *cohen_kappa_score* function from the *sklearn* package. Accuracy was also calculated between researchers and the sentiment calculator using the *accuracy_score* function from the same package. Furthermore, the answers that were classified differently by the raters and by the algorithm were manually analysed individually, and words were added to the dictionary. For example, the word "*seguro*" ("*safe*") was added to the dictionary. In this way, the dictionary can be used for future work in which patient experience in digital healthcare is analysed.

Furthermore, direct answers to interview questions were manually selected from the dataset. In order to make conclusions about specific topics, TFIDF scores unigrams and bigrams were calculated for each question. Finally, the sentiment associated with direct answers to each question was calculated by using the sentiment calculator.

4.2.3.2.2 ML approach

The second method used for sentiment analysis was the ML approach. The dataset employed to train the model was the manually labelled one, previously used to calculate the accuracy in the dictionary-based approach. Firstly, the dataset was examined by creating a bar graph to visualize the dimension of each class: positive, negative, and neutral. After that, TFIDF was used to vectorize the data. The data was then split into training set and test set, in a ratio of 25% for the test set and the remaining for the training one.

The chosen classifiers were NB and SVM. This choice was based on the results of a study performed by Khanbhai et al[5].; The authors systematically searched for articles examining NLP to analyse free-text patient feedback and reported that the performance metrics are better with these models and therefore, they are the most used supervised algorithm and a baseline method for text categorization[115,135]. Also, specifically the NB classifier is known to be faster compared to other classifiers and works well with small amounts of training data[8].

To improve the model's performance, the dimensions of the classes in the dataset were balanced. With that purpose, we followed Sarakit *et al.*[136] proposed framework using the SMOTE algorithm, in which the authors oversampled the minority class until it equalled the majority class in order to obtain a balanced dataset. In this case, the *SMOTE* function from the *imbalanced learn* package was applied to the dataset. The parameter *sampling_strategy* was set to *'minority'* so that the algorithm only resampled the minority class. Then, the classifiers were applied once again and the results were compared with previously obtained results.

4.2.3.3 Topic modelling

To create the unsupervised topic models, the primary Python libraries used were *gensim*, *pyLDAvis* and *sklearn*. Firstly, clean and lemmatized text was used to build a bag of words corpus using the *genism* library. Furthermore, a dictionary was created with the same library, and a limit was set to remove words that appeared in less than 5 documents and in more than 80% of the documents. The resulting corpus consists of a mapping with words and their identification numbers.

Regarding the alfa hyperparameter typical default value for the alpha is 50 divided by the number of topics. A lower alpha leads to a higher concentration of topic distributions within documents, meaning that documents score high on a few topics rather than low on many[137]. After trying various values for alfa, the one that showed the best results was $5/k$, k being the number of topics, instead of $50/k$ as suggested in the literature.

After that, a function was created to iterate over various numbers of topics, starting with two topics, until twelve topics, two in two. Then, for each number of topics, the function calculates the semantic coherence within that specific number of topics. With this function, a coherence vs. number of topics was plotted just like Allahyari *et al.*[138]. The number of topics chosen was the one with highest topic coherence values. Finally, using the $pyLDAvis$ package, a diagram was generated with the output topics and the top 30 terms in each topic, shown in appendix 1. Finally, a table was created with the top terms for each topic. This table only contains the words that were

considered significant for the study. Words such as "*algum*" and "*coisa*" (in english "any" and "thing") were excluded.

Posteriorly, each topic was organized into a list and used as an input in the sentiment calculator. Each topic in was represented in a colour that represents the underlying sentiment obtained by the calculator.

Chapter 5

5 Results and Discussion

The following chapter discusses various results from the NLP of interview video transcriptions. In order to make this section understandable for non-Portuguese readers, every term in Portuguese language is followed by its translation in English.

5.1 Collecting structured data to evaluate patient's experience with digital healthcare services

5.1.1 Questionnaire to evaluate patient's experience with digital health services

Table 5.1 shows in bold the items that were chosen for translation and used in the final survey. The final survey consists of 13 items containing questions from each topic. The final survey is available at Appendix 2 and is currently being used in this project and on other ongoing studies.

Table 5.1 - Questions from TUQ, RMSS, and SUTAQ organized by topics. Questions selected for final questionnaire are represented in bold.

Service quality	• **[*The equipment*] has been explained to me sufficiently**[101] • **If technical problems occur, the staff are quick to respond and fix technical problems**[108] • **I am satisfied with [*the service*]**[101,108] • I would recommend [*the service*] to other patients[108] • **[*The equipment*] should be recommended to people in similar condition to mine**[101] • I would use [*the service*] again[100] • Overall, I am satisfied with [*the service*][100]
Interaction with medical provider	• The staff are responsive to my questions and concerns[108] • I am satisfied with the amount of communication I receive from the staff.[108] • **I am satisfied with the quality of my interaction with the staff.**[108] • I can easily talk to the clinician clearly using [*the service*][100] • I was able to express myself effectively[100] • Using [*the equipment*] I can see the clinician as well as if we met in person[100] • I feel comfortable communicating with the clinician using [*the equipment*][101]
Quality of care	• [*The service*] provides for my healthcare need[100] • **I think the visits provided over [*the equipment*] are the same as in-person visits**[100] • [*The equipment*] I received has helped me to improve my health[101] • [*The equipment*] has allowed me to be less concerned about my health and/or social care[101] • [*The equipment*] can be a replacement for my regular health or social care[101] • [*The equipment*] can certainly be a good addition to my regular health or social care[101] • [*The equipment*] is not as suitable as regular face to face consultation with the people looking after me.[101] • [*The service*] interferes with the continuity of the care I receive (*i.e.* I do not see the same care professional each time)[101] • [*The service*] has allowed me to be less concerned about my health status[101]
Self management	• [*The equipment*] has made me more actively involved in my health[101]
Privacy	• **[*The equipment*] I received has invaded my privacy**[101] • [*the equipment*] makes me worried about the confidentiality of the private information being exchanged through it[101]
Interaction with the equipment	• [*The equipment*] is easy to use[108] • It was simple to use [*the equipment*][100]

	• **[*The equipment*] is reliable and has few technical problems**[108] • [*The equipment*] can be trusted to work appropriately[101] • **It was easy to learn how to use [*the equipment*]**[100] • The way I interact with [*the equipment*] is pleasant[100] • I like using [*the equipment*][100] • [*The equipment*] is simple and easy to understand[100] • **Whenever I made a mistake using [*the equipment*], I could recover easily and quickly**[100] • **[*The equipment*] made me feel uncomfortable, e.g. physically or emotionally**[101]
Easy to fit into daily routine	• **The amount of time it takes to complete my daily home [*tasks*] is acceptable**[100,108] • [*The service*] saves me time travelling to a hospital or clinic[100] • [*The equipment*] I received saved me time in that I did not have to visit my GP clinic or other health/social care professional as often[101] • **[*The equipment*] I received has interfered with my everyday routine**[101]

Items from this questionnaire were collected from TUQ, TSQ, SUTAQ, TSUQ, PSSUQ, and RMSS. Posteriorly, from the six surveys mentioned above, only three were selected (TUQ, SUTAQ, and RMSS), as the remaining three had very similar questions. The five-point Likert scale was chosen as its most often used in these questionnaires and a universal method of collecting data. We considered this scale the most simple and intuitive. It covers all degrees of agreement (1 for "strongly disagree", 2 for "disagree", 3 for "neutral", 4 for "agree", and 5 for "strongly agree")[109].

The final questionnaire has already been applied to 10 participants of another project. Similarly to the *CardioFollow.AI* project, this project involves cardio-thoracic surgical patients using a digital device. The researchers stated that, in general, the items didn't raise doubts and the feedback received by the researchers is significantly better than the feedback obtained from the application of the UEQ questionnaire. However, patients demonstrated difficulty understanding the statement "*This service interferes with my daily routine*"; namely, the word "*interferes*" seems hard to interpret.

Validating a questionnaire is a complex methodology that requires a large-scale pilot test. Nonetheless, initially, to assess the extent to which the survey items are representative of the theoretical construct it's meant to study, a panel of experts who are familiar with digital healthcare services should be tasked with evaluating the content validity of the questionnaire. Posteriorly, a preliminary pilot study with a group of 30 participants is highly advisable to evaluate the feasibility, that is, to understand if there is confusion in any questions such as the one mentioned above[86]. Evaluating feasibility is particularly relevant in a healthcare context as there is a higher probability that some respondents have low literacy.

Additionally, this step will allow us to understand if 13 items are sufficient to measure the construct of interest or if it is too extended, causing respondents to experience fatigue or lose motivation in completing the questionnaire. For example, if all participants answer one of the items with top scores, the measurement scale may have to be modified to include more options that distinguish the high values. Another possible source of error in this questionnaire is its translation, as questions may be interpreted differently before and after translation. Therefore, this preliminary pilot study can also be helpful to ensure that the translated version of the questionnaire is equal to the original version by asking the participants to say what they think each item means[86,87]. Finally, we can also get a rough idea of the response distribution to each item to assess whether there is enough variation in response to justify moving forward with a large-scale pilot test.

Posteriorly, a large-scale pilot test is necessary to evaluate if questions that measure the same topic are consistent; inter-content reliability must be calculated within each topic. Reliability is a metric that depends on the sample of respondents. Therefore, it should be estimated each time this

questionnaire is administered. Furthermore, we should administer this survey with similar pre-existing validated instruments to the same group of individuals. With the answers to both questionnaires, correlation measures are used to compare different constructs of the same measure[86].

5.2 Analysis of unstructured verbal data using NLP

5.2.1 Pre-processing and Exploratory Analysis

Figure 5.1 shows the text data before and after it was pre-processed in python. Regarding the lemmatizer, as we can see, it works well in some cases, but in general its performance is still not very accurate, due to the lack of available corpus for Portuguese. For example, in the case of Figure 5.1, we can see that the word "*sinto*" ("*feel*" in the singular present tense) should be replaced by its respective lemma "*sentir*" ("*feel*" in its standard verbal form), and in fact such change does not occur.

INPUT	OUTPUT
<Researcher_1>: E foi fácil para si utilizar? **<Patient>:** Foi, foi fácil, a recuperação foi (...) foi, e está a ser ótima, sinto-me diariamente com (...) com mais força, sinto-me com mais força (...)	**['fácil', 'recuperação', 'ser', 'ótima', 'sinto', 'diariamente', 'força', 'sinto', 'força']**

Figure 5.1 – Portion of raw text (left) and the exact text after removing researchers' questions, tokenizing, punctuation removal, stopword removal, and lemmatizing the tokens (right)

The most frequent unigrams, bigrams and trigrams calculated with the *FreqDist* functions were not meaningful and weren't very useful to for exploratory analysis purposes. Examples of the most common words in the dataset are "*aqui*" ("*here*") and "*coisa*" ("*thing*"). Nevertheless, one could infer that the overall satisfaction was good, as the unigram '*bem'* ("*good*" or "*well*") was the most frequent unigram, and that there are no apparent negative words in any of the n-grams. The trigrams are more meaningful, as they already include words like "*recuperação*" ("*recovery*"), and "*serviço*" ("*service*"), however, the occurrence frequency of each trigram is too low to be considered significant to characterize the overall dataset, occurring three or four times in the text.

Figure 5.2 shows the most frequent adjectives, verbs and nouns present in the document obtained using the *FreqDist* function from NLTK. Regarding the adjectives, we can see that the most of them are positive, such as "*bom*" ("*good*"), "*importante*" ("*important'*'), "*interessante*" ("*interesting*") and "*excitante*" ("*exciting*"). Therefore, extracting the adjectives is a good alternative for a simple frequency analysis, namely in patient experience datasets, and may give more insights of the sentiment tone of the dataset in its whole. In terms of the POS tagger's performance in Portuguese, I identified some mistakes, for example, the word "*hospital*" ("*hospital*") was identified as an adjective when it should be tagged as a noun.

Another drawback is that semantic aspects are not considered, which generates highly inaccurate results. For example, the word "*preciso*" was identified as a frequent adjective. However, this word can have two meanings: the verb "*need*" in the simple present tense, or the adjective "*precise*". After reading the interview transcripts, we knew beforehand that this word was

mostly used as a verb and not as an adjective. This problem also applies to the word "*bom*", which is identified as an adjective but can be used as an interjection. This aspect could be improved by taking in consideration the words surrounding it, and in this case, we would know the word is an adjective if it is associated to a noun. *spaCy* provides a function for this purpose, called the "dependency parser," that learns sentence segmentations and identifies syntactic dependency of words within a sentence. Using this function could be implemented in future work and improve the results.

On the other hand, verbs and nouns don't show to have much significance. The most common verbs identified were "*dizer*" ("*say*") and "*fazer*" ("*do*"), which in my view is coherent with the fact that this interview was part of a telemonitoring service, and may be due to the fact that patients had to measure their vitals ("*fazer medições*") and speak to the medical staff during their recovery period. However, looking at raw verbs and nouns solely alone may not lead to any conclusions about the dataset.

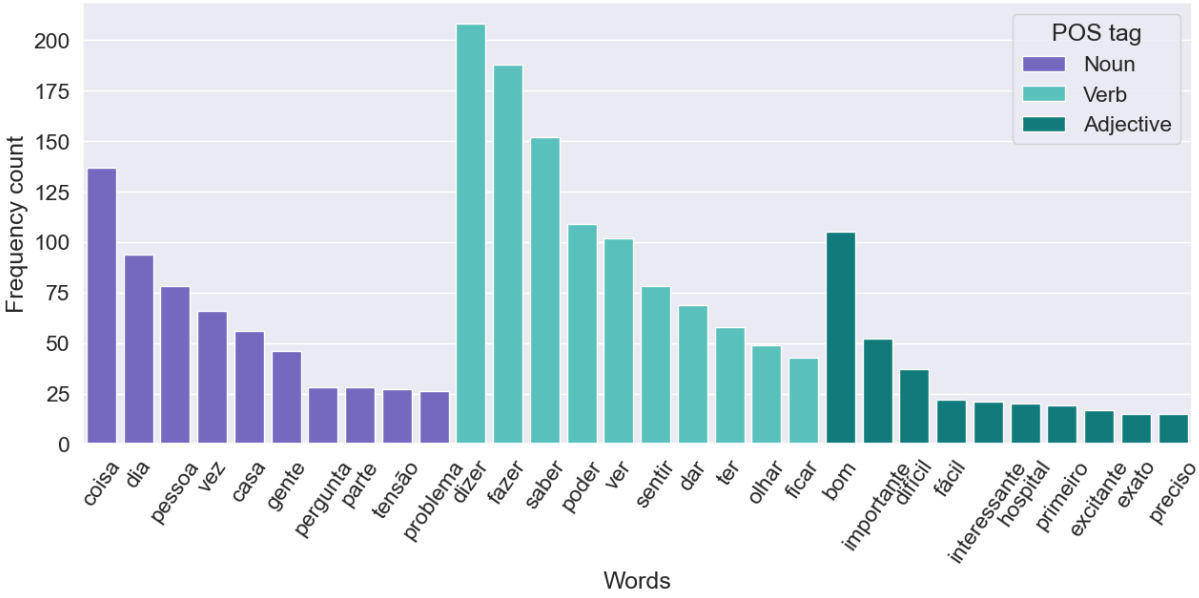


Figure 5.2 - Ten most common nouns (purple), verbs (blue) and adjectives (green) present in the dataset.

5.2.2 TFIDF vectorizer

The output of the TFIDF vectorizer was a document-term matrix containing the vocabulary and the IDF score for each feature, that is, for each unique term. A 20 by 1622 words sparse matrix was obtained, in other words, the dataset contains a total of 1622 unique words. Figure 5.4 and shows the n-grams with highest TFIDF scores. Figure 5.5 shows the highest TFIDF scores within each document, and Figure 5.3 shows a word cloud to help visualize each word importance depending on its score. In this heat map, it is possible to visualize the main topics to which each patient gave more importance. Figure 5.5 may be an interesting visual representation for patient or costumer experience analyses in which each individual opinion is worth analysing.

The obtained terms give much better insights into the topics of the text compared to a simple frequency count distribution. By reading the unigrams "*projeto*" ("*project*") and "*tecnologia*", ("*technology*"), one could easily infer that the main topic of the document is a technological project for monitorization. The same thing applies to the trigrams "*Lisboa, ser, monitorizar*" ("*Lisbon, be,*

monitorized") and "*através, projeto, fiquei*", ("*through, project, stayed*"). Therefore, the algorithm shows results that are consistent with the reality of the study. The bigrams didn't give any insights on the study or patient experience.

Regarding patient experience, the words "*positivo*" ("*positive*"), "*valer, pena*" (from "*valer a pena*", Portuguese expression which can be translated to "*worth it*"), "*ajudar todo dia*" ("*help, all, day*"), already suggest that the main satisfaction feedback is positive. Likewise, several words mention disease related terms such as "*dor, sentir, peito*" ("*pain, feel, chest*").

The fact that these terms obtained the highest scores indicates that the most mentioned health aspects are chest pain, which is consistent with the study since they are cardiac patients. In addition, the unigram "*gordura*" ("*fat*") was classified with the highest score. This result remains consistent with the study, as some patients mentioned their physical condition or eating habits. Also, the terms "*carregar*" ("*charge*") and "*pilha*" ("*battery*") appear with high TFIDF scores. These results allow us to conclude that it is worth analysing aspects related to the battery charge of some of the equipment in the telemonitoring kit. Furthermore, as there was no question about this specific subject, it demonstrates that it was an important aspect for patients.

Figure 5.3 - Word cloud of the terms with highest TFIDF scores. Larger words have higher scores.

Unigrams	Score	Bigrams	Score	Trigrams	Score
gordura	0.496	viver, Lisboa	0.280	ir, bem, costa	0.158
projeto	0.385	gordura, gordura	0.236	dor, sentir, peito	0.142
pessoal	0.353	ali, tudo	0.199	Lisboa, ser, monitorizar	0.141
carregar	0.320	em este, caso	0.190	através, projeto, fiquei	0.141
nada	0.318	valer, pena	0.170	algum, sinto, necessidade	0.141
tecnologia	0.315	quatro, mil	0.160	algo, acontecer, automaticamente	0.141
viver	0.311	ir, ter	0.160	ajudar, viver, Lisboa	0.141
positivo	0.309	ir, estranhar	0.159	ajudar, todo, dia	0.141
eu	0.302	ser, oferecer	0.157	ajudar, precisos, viver	0.141
pilha	0.297	bem, costa	0.157	acontecer, automaticamente, avisar	0.141

Figure 5.4 - TFIDF scores for the highest scored unigrams, bigrams, and trigrams

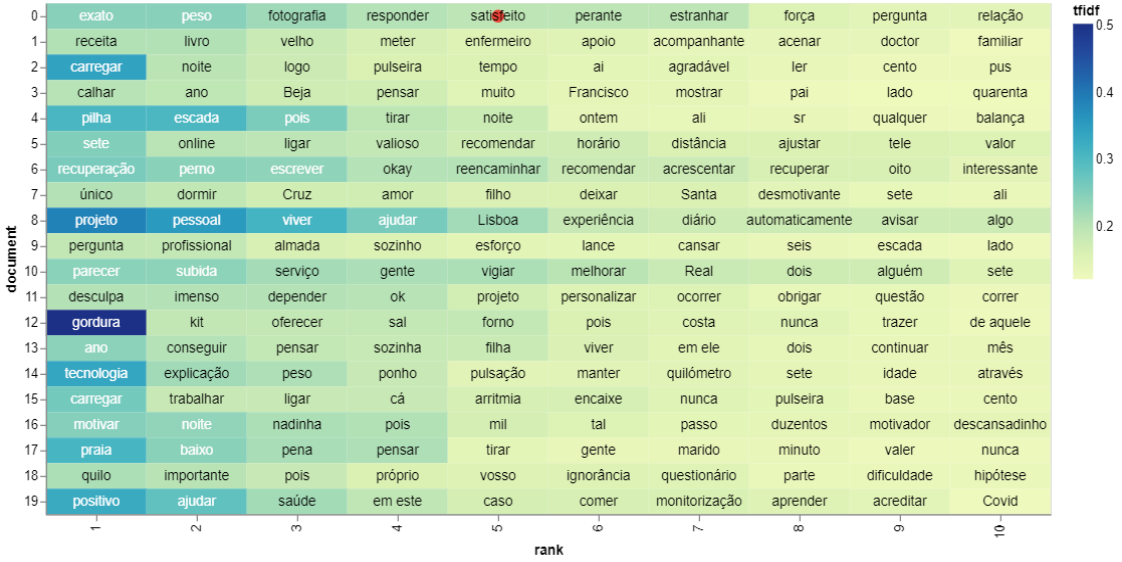


Figure 5.5 - Terms with highest TFIDF scores in each document

5.2.3 Sentiment Analysis

5.2.3.1 Dictionary-based classifier

The dataset consisted of 1597 tokenized, clean and lemmatized patient answers. After two independent raters labelled each answer, the inter-rater coherence score obtained was $k = 0.454$. Cohen suggested that this value of coherence score is considered "moderate"[139]. A likely reason for this low value is that one rater knew the dataset's context, and the other didn't. Therefore, the first rater could infer an answer's sentiment based on previous knowledge. On the other hand, it may also suggest that the patient's responses in the dataset didn't have a clear sentiment associated. Specifically, the interview parts when the patient was not directly answering a question, naming a sentiment was not an obvious task. Also, a portion of the answers had a mixed sentiment, that is, having positive and negative words.

The accuracy of the computer, compared with the coded data, was $k = 0.680$. This value is mainly because this method consists of a dictionary-based approach, which doesn't consider the semantics of a sentence but the meanings of individual words. Furthermore, the dictionary provided for the calculator may need to be more representative of each sentiment, especially within the scope of healthcare. However, the obtained accuracy value can be a matter of the dataset itself. As suggested by the interrater score, the dataset is not obvious to classify as "positive", "neutral", or "negative".

The obtained accuracy was reasonable but one must recall that the words used in a healthcare context expressing positive and negative sentiments differ from others. For example, the term "safe" isn't considered positive sentiment words in the used dictionary but express positive experience in healthcare. Another possible reason for low accuracy is the fact that this calculator only considers the number of positive or negative words. Therefore, if a comment contains multiple sentences with positive and negative sentiments, then it may be considered neutral. For example, consider a comment that contains two sentences. One of the sentences contains a negative word and is scored -1 by the calculator. The other is scored +1 as it contains a positive word. The comment would be considered as neutral as the sum of both scores is zero.

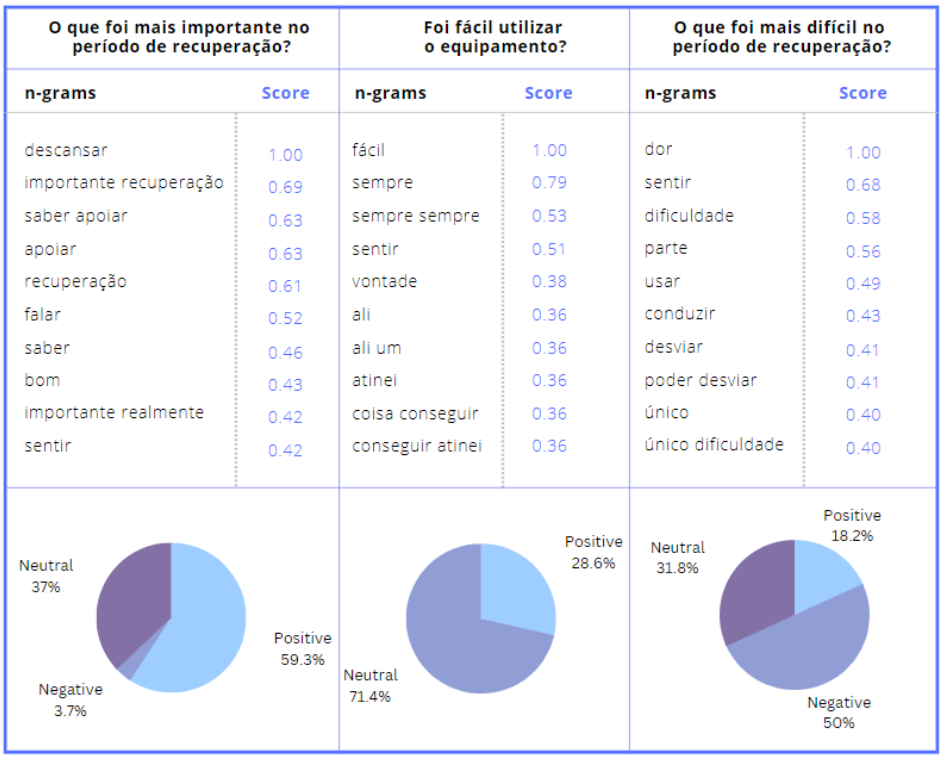


Figure 5.6 - Unigrams and bigrams with highest TFIDF scores for three interview questions: (Left)"What was most important to you during the recovery period?", (Middle) "Was it easy to use the equipment?", (Right) "What was hardest during the recovery period?". Sentiment percentages are represented in pie charts for each question.

Figure 5.6 shows the highest TFIDF scores for three interview questions. Regarding the first question, "What was most important to you during the recovery period?", we can see that most words relate to the support patients received. Even though it is not possible to understand from who or what it is referring, the terms "*apoiar*" ("*to support*") and "*falar*" ("*to talk*") indicate that the support patients receive is what is most relevant in a recovery period. This answer was expected, as it relates to one of the identified dimensions contributing to the digital healthcare experience identified in chapter 3.2.1: support and availability. Indeed, we knew beforehand that patients mostly talked about family support and the support they received from the telemonitoring service when answering this question.

Also, the word "*descansar*" ("*to rest*") appears as most relevant, which is aligned with the fact that patients feel more "in peace" with themselves knowing they are being monitored during their recovery period. The main sentiment obtained for this question was positive. However, this sentiment isn't relevant to analyze for this specific question because the question is already inherent to positive sentiment. Similarly, the question "What was hardest for you during the recovery period?" is inherently associated with a negative sentiment. This is aligned with the sentiment score obtained, as 50% of the answers received were negative.

The word "*dor*" ("*pain*") was the most relevant when patients were asked to discuss the most challenging part of the recovery. The results are aligned with the fact that, based on the pre - knowledge of the dataset, patients only had negative comments about the discomfort during the recovery period. The remaining n-grams don't allow us to make significant conclusions about the question answers. Finally, the question "*What is easy to use the equipment?*" had as most relevant

words the word "*fácil*" ("*easy*"), "*sempre*" ("*always*"), "*conseguir*" ("*to accomplish*") and "*atinei*" (Portuguese expression that can be translated to "*I figured out*"). Additionally, none of the answers were classified as negative by the calculator. Therefore, we can conclude that it was generally easy for patients to learn how to use the equipment.

5.2.3.2 Machine Learning approach

Figure 5.6 (left) shows the size of the labelled dataset's positive, negative, and neutral classes used to train the ML model. As shown, the classes are imbalanced, the negative class being the minority class and the neutral the majority. Having previous knowledge of the dataset, we could preview that this dataset would be imbalanced because the patients gave little negative feedback when interviewed. Furthermore, the minor negative feedback obtained was related to the disease and not the service quality. Therefore, a more balanced dataset would be needed to improve the machine-learning models' performance.

The accuracy value for the NB model was $acc = 0.74$. In fact, the model was mostly correct when classifying the majority class data, correctly classifying 90% of the neutral class responses. This value causes the accuracy value to be high and disregards the results obtained in the minority class. Hence, accuracy is an inadequate measure to quantify predictive performance in an imbalanced setting and can be misleading in these cases. In cases of imbalanced data, the value of f1 is more useful because it considers the measure of precision and recall. The obtained f1 for the neutral class was 0.84 and 0.11 for the negative class.

After applying the SMOTE technique, classes were balanced; each class had 784 examples. The obtained results, represented in Figure 5.7 (middle and right), were significantly better after oversampling. However, as shown in table 5.2 the precision values remain higher for the neutral sentiment class and lower for the negative sentiment class. 30% of the correctly evaluated cases were of negative sentiment, and 82% were of neutral sentiment. Since this study aims to create a tool that supports assessing patients' experience, it is crucial that the model classifies cases of negative sentiment as such. Thus, more than the precision value, we are interested in evaluating the recall value, since it is essential to assess false negatives, that is, comments on the patient's experience that have a negative feeling and are classified as positive. The recall value for the negative sentiment class is 0.62, that is, 62% of the negative sentiment cases were classified as such. Comparing both models, despite the SVM model having higher accuracy, it has a recall for negative cases of 0.54, which is lower than the previous model. As such, the second model performs better for the neutral instances but not for the other cases.

It should be noted that, in this case, over-sampling was performed; the algorithm generated new cases in the data to increase their effect on the classifier while all instances of the majority class were maintained. This approach can lead to overfitting, introducing duplicate instances into an already reduced database. SMOTE can also allow overgeneralization, that is, mixing classes, since it generalizes the negative class without considering the majority class.

Compared with Dey et al[135]., which used a similar approach to predict movie review sentiments, the results were significantly poorer. The accuracy they have obtained for 2000 movie reviews was 79.14%, with a recall of 77.26% for positive cases and 81.02% for negative. The performance achieved in this study may be due to the size of the dataset, the fact that each answer doesn't have a clear sentiment, and the dataset only contains 1597 patient answers with very few negative examples. Other possible reasons are related to text pre-processing, as the available options for the Portuguese language are scarce. Comparing this model with the dictionary-based one, this

approach is preferable as it is trained with data from the same context it will be applied to, making it more adequate.

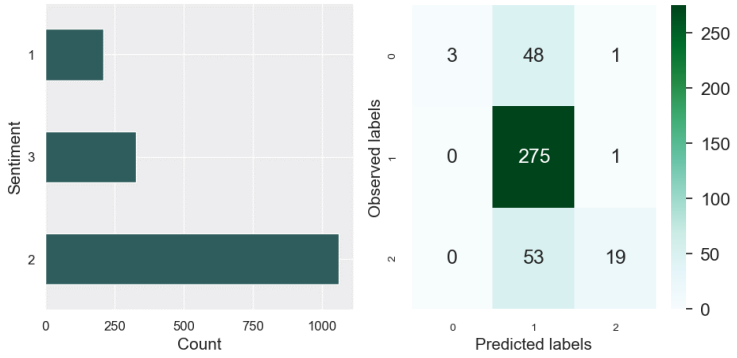


Figure 5.7 - (Left) Class Dimension. 1 is negative, 2 is neutral and 3 is positive sentiment. (Right) Confusion matrix for the NB model applied to the labelled sentiment data

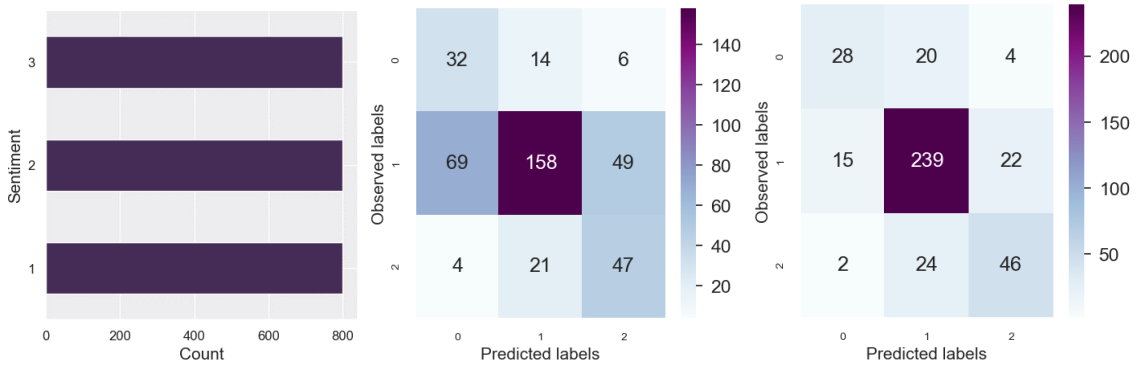


Figure 5.8 - (Left) Class dimension after applying the SMOTE algorithm. (Middle) Confusion matrix for the NB Model trained with the oversampled dataset. (Right) Confusion matrix for the SVM model trained with the oversampled dataset

Table 5.2 - Precision, Recall an F1-score for the NB (blue) and SVM (green) classification models

		Precision	**Recall**	**F1-score**
NB	**Positive**	0.30	0.62	0.41
	Neutral	0.82	0.57	0.67
	Negative	0.46	0.65	0.54
	Accuracy	0.59		
SVM	**Positive**	0.62	0.54	0.58
	Neutral	0.84	0.87	0.86
	Negative	0.64	0.64	0.64
	Accuracy	0.78		

5.2.4 Topic modelling

LDA utilizes optimization algorithms that require the user to select the number of features desired before developing a model based on the vocabulary of the corpus. In this study, the number

of topics was used based on the coherence score, as mentioned previously. The number of chosen topics was 6, with a coherence value of 0.2274, which was the most significant value obtained, as shown in Figure 5.8. However, one should note that LDA is a generative probabilistic model, using Dirichlet distributions, which depends on random number generators. Therefore, every time the code is executed, different coherence results are obtained. Still, it is to emphasize that this method was only used to make an initial selection of the number of topics and that interpretability of topics is a more important criterion for social science purposes. The goal is to describe the data with less dimensions to make its interpretation simpler, but with enough dimensions so that there is no loss of information[137]. Therefore, in these types of studies, it is natural that the researcher manually inspects these topic models and, decides on the correct number of topics by trial and error, as long with interpretation the words in the topic.

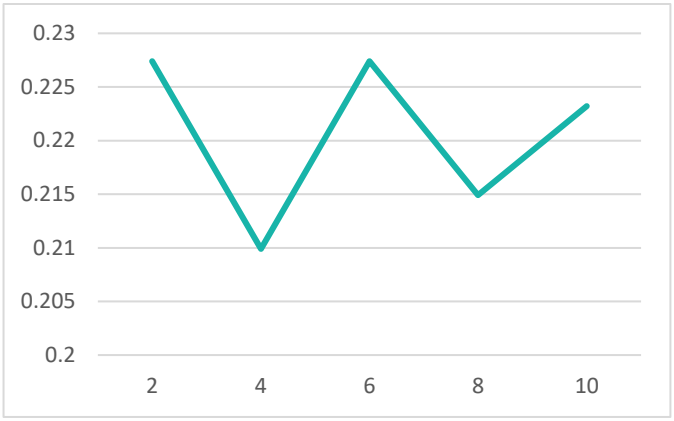


Figure 5.9 - Coherence values for each number of topics in the LDA model

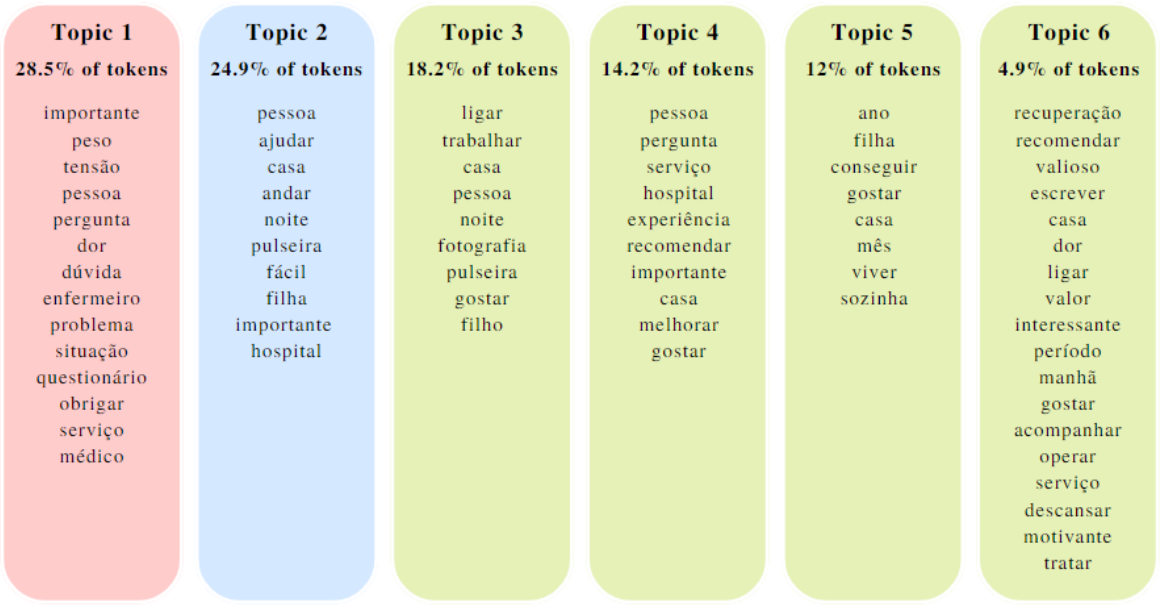


Figure 5.10 - Topics generated by the LDA model. Each topic is represented by a colour that represents the obtained sentiment by the sentiment calculator. Red is negative, blue is neutral, and green is positive

Some studies that explore NLP in patient experience use databases like Press Ganey, in which comments are already classified into topics[4]. The advantage of using topic modelling strategies like LDA is that the research isn't limited to previously known topics but allows the model to discover unexpected topics. This fact is even more relevant in healthcare context. Methods that are sensitive to new topics in patients' experiences are essential in the constantly changing healthcare environment[1].

By analysing Figure 5.9 , it is easily noticeable that there is little coherence within each topic, that is, it is difficult to infer the topic by reading each topic's words. Although, in general, there is a reduced coherence between the words of each topic, it is possible to infer the theme of topics 1 and 6. In topic 1, we can see that it mainly includes words related to recovery from surgery and aspects disease-related (*i.e.* "*peso*", "*tensão*", "*dor*", "*enfermeiro*", "*médico*", in English, "weight", "pressure", "pain", "nurse", "doctor"), together with words related to the hospital's support when patients raise concerns about their recovery (*i.e.* "*pergunta*", "*dúvida*", "*problema*", "*situação*", in English "question", "doubt", "problem", "situation ").

Topic 6 includes the smallest percentage of tokens. Still, it is easy to see that it is related to the telemonitoring service due to the use of the words "*acompanhar*", "*recuperação*", and "*serviço*" (in English, "accompany", "recovery", and "service"). Additionally, this topic includes words that reveal the positive feedback received from patients regarding the service (*i.e.* "*recomendar*", "*valioso*", "*valor*", "*interessante*", "*gostar*", "*motivação*", in English, "recommend", "valuable", "value", "interesting", "like", "motivating"). The positive feedback was confirmed by the sentiment calculator, which output the result "positive".

Topic 2 consists mainly of words related to recovery at home (*i.e.* "*casa*", "*noite*", in English "home", "night") with reference to family members, in this case the word that stands out is "*filha*" (in English, "daughter"), and the telemonitoring system, with reference to the word "*pulseira*" and "*hospital*" (in English, "bracelet" and "hospital"). Note that the word "*pulseira*" in this context refers to the smartwatch that was part of the telemonitoring kit. It was not possible to extract a theme for topics 3, 4, and 5, but they are more related to topic 2, that is, with recovery at home. The underlying sentiment was positive because of the word "gostar".

Regarding the sentiment of each topic, we can conclude that most of the topics are positive and only one is negative, the one related to the surgery recovery. Therefore, with the topic modelling approach, the major conclusion is that that the overall patient sentiment regarding the *CardioFollow.AI* service is positive. Additionally, without knowing about the topics of the interview, researchers can conclude that the main concerns patients have are related to doubts about their recovery and the pain associated with recovery, by analysing topic 1. Also, general satisfaction was good as most topics are positive and contain mostly positive adjectives. To conclude, even if this approach doesn't immediately answer the research question, it can be used for preliminary analysis and enhance further manual analysis by inductively showing which topics occur. Consequently, LDA may give insights to the researcher to create or improve the codebook that would be used initially, and suggesting new codes.

5.3 Limitations

This study's main limitation is related to the dataset's size and quality. The dataset contained only 20 interview transcripts, and only patient responses were considered, some of which were relatively short. NLP and ML models are not very effective on small datasets as these algorithms rely on frequency counting, word and n-gram repetition, and patterns. As such, the models may be less fruitful due to the reduced amount of raw text data[4]. One should also consider that the dataset consists

of a conversation transcription; therefore, there are speech expressions that wouldn't appear in written text data or scientific articles, for example. All these aspects increase the noise and complexity of the data. In addition, the videos that made up the dataset not only consisted of the interview but also included recordings of the conversations before, during, and after the interview that were unrelated to the subject. Additionally, recordings included the moment the patients filled out a quantitative satisfaction questionnaire.

A possible approach to solve this would be only to select the relevant parts of transcripts. However, this was not done because, in addition to being a very time-consuming job, it could increase the bias for an investigator to choose what is considered to be relevant. Nevertheless, considering this limitation, the main contribution of this study lies in its methodology rather than its performance.

As mentioned by Cammel *et al.*[9], a notable aspect that interferes with these results, is that this dataset was obtained from patients' answers to a semi-structured interview. The topics in the document are influenced by the questions asked. Regarding the sentiment analysis, the sentiment of an answer was related to the question asked. For instance, the question "*what was most difficult during the recovery period?*" is associated with negative sentiments.

Regarding topic modelling, considering the obtained results and my literature review, it is not recommended using this approach for short text. Other than the LDA model, LSA and NMF were also implemented, but the results weren't worth analysing because of how inaccurate they were. Particularly for LDA, which infers topics iterating through documents, in case there aren't enough words, it is difficult to infer a reliable topic distribution for that document. As observed in the literature review of this research document (section 2), topic modelling is mostly used in big data from social media platforms or customer review datasets, for example[137]. However, even in these cases, if we consider, for example, a twitter dataset, each post or comment on twitter is typically short text, and a post may not discuss one single topic. Similarly, in the case of this study, the dataset consists of relatively short answers for interview questions, they don't discuss a specific topic coherently.

There are some limitations with the pre-processing of textual data due to the lack of resources available for the Portuguese language. For example, lemmatization is important because the computer cannot distinguish the word "eat" from "ate" even though they have the same meaning. However, the function that exists in python for lemmatization in Portuguese still has many errors, probably due to the available dataset in Brazilian Portuguese.

The use of frequencies can provide useful information; however the model treats each term without considering its context. Evidently, depending on the context, the meaning of words and interpretations can vary greatly. This situation applies not only to the word count but also to each individual response. Currently, a model that can evaluate text considering the context is BERT[140], an algorithm that is the ideal alternative for healthcare institutions, due to its ability to process large amounts of text and its processing speed.

6 Conclusion

This study describes the development and validation of a comprehensive tool for surveying patient experience in digital care in Portuguese language. It contributes to the literature as there are currently no surveys for patient experience in digital healthcare in Portuguese. Furthermore, even though the survey wasn't validated, we gathered validated questionnaire items from the literature adequate for digital healthcare.

This work also provides a tool that may be useful for researchers to analyse patient experience in healthcare services in the Portuguese language, replacing the tedious tasks of reviewing textual data. It consists of a comprehensive method that combines an NLP pipeline and a visualization and supports healthcare professionals in defining quality improvements from the results.

By combining fundamental numerical analysis of textual content from patient feedback and ML, this study demonstrated that patient comments in the Portuguese language could be processed to obtain meaningful and actionable insights. Firstly, NLP enhances the analysis of patient feedback by discovering the most relevant words in documents using TFIDF statistics, which is a better measure than solely calculating word frequency. Besides, obtaining n-gram frequency is helpful to bring light to aspects that may not be clear by analysing single words. POS tagging has shown to be beneficial, especially when extracting adjectives; however, the libraries that perform these tasks in Portuguese could be more accurate.

Training NB and Support Vector ML models with labelled data proved to be a valuable approach to classifying documents according to their sentiment at high speed. However, the obtained accuracy results were poor; a more extensive dataset would enhance this performance. The same limitation applies to topic modelling; yet, although the performance wasn't ideal due to the reduced dataset, the combination of topic modelling with sentiment analysis showed to be a good indicator of the aspects to work on and aspects to celebrate when analysing patient experience. Also, topic modelling is extremely beneficial because it doesn’t limit information to specific topics and lets the model discover unexpected issues. The pipeline allowed us to conclude that most of the negative comments were related to disease and recovery aspects and that the overall satisfaction with the *CardioFollow.AI* project was positive.

In conclusion, this work contributed to developing tools to collect patients' experiences in European Portuguese. We proposed a questionnaire that makes it simple to analyse patient experience in digital healthcare services and automatic tools to allow the collection of experience through interviews with patients in the European Portuguese language. Simple questionnaires and unstructured data analysis are relevant, considering that patients have low literacy and are not prepared to report their experience in complex data surveys. The results will be used as outcomes of the *Cardiofollow.AI* project, as patient experience is a relevant indicator of value in digital healthcare services.

6.1 Future Work

Firstly, the proposed questionnaire to evaluate patient experience using digital healthcare services must be validated and tested for feasibility with a small population sample, and then tested for validity and reliability with a larger population pilot test.

A possible approach to increase the sentiments dictionary's accuracy is to create a more complete and adequate dictionary. To complete the dictionary list, a researcher could read through all sentences

that weren't classified correctly and select the words that make that sentence positive or negative, adding them to the list. The downside of this process is that manual labelling is required to test the calculators' accuracy and to review all comments. Furthermore, to obtain more meaningful results, researchers could manually label each word in the dictionary with a score according to how positive or negative the word is; for example, the term "*bad*" is negative but would have a lower score than the word "*terrible*". Posteriorly, the calculator could be used to score each answer in the dataset by calculating an average of the scores of each word.

Regarding the topic model, the accuracy could have been computed by manually annotating patients' answers and comparing them to the model's results. To do so, one must select part of the dataset and assign each response to one of the generated topics. Another interesting approach made by Doing-Harris *et al.*[115] which could be future work in this thesis, is to select only negative sentiment data and use them for topic modelling to identify the aspects causing a negative experience. Ultimately, to obtain more meaningful results from the topic modelling approach, instead of analysing a whole transcript at once, researchers could explore the answers to each question separately. In this way, it would be easier could understand the underlying topics that are referred to, within each answer. This is already done, in section 5.2.3.1, by using the sentiment calculator for each interview question separately.

Finally, a larger patient experience dataset is needed to examine the feasibility of the model. The main goal for future work is to optimize the code in order to make it generalizable to any dataset. In this way, by inputting transcript files, the model would return a visual representation of the results, allowing researchers to conclude about patients' experience in a healthcare study in a comprehensive way.

7 References

1. van Buchem MM, Neve OM, Kant IMJ, Steyerberg EW, Boosman H, Hensen EF. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM). *BMC Med Inform Decis Mak*. 2022;22(1). doi:10.1186/S12911-022-01923-5

2. Ish D, Parker A, Osoba O, *et al. Using Natural Language Processing to Code Patient Experience Narratives: Capabilities and Challenges*. RAND Corporation; 2020. doi:10.7249/RRA628-1

3. Beattie M, Murphy DJ, Atherton I, Lauder W. Instruments to measure patient experience of healthcare quality in hospitals: A systematic review. *Syst Rev*. 2015;4(1):1-21. doi:10.1186/S13643-015-0089-0/TABLES/6

4. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. *Appl Clin Inform*. 2020;11(2):242-252. doi:10.1055/S-0040-1708049

5. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform*. 2021;28(1):100262. doi:10.1136/BMJHCI-2020-100262

6. Riebling NB, Norouzzadeh S, Reeder G, *et al.* Quantifying patient satisfaction with process metrics using a weighted bundle approach. *BMJ Open Qual*. 2019;8(1):458. doi:10.1136/BMJOQ-2018-000458

7. Gualandi R, Masella C, Viglione D, Tartaglini D. Exploring the hospital patient journey: What does the patient experience? *PLoS One*. 2019;14(12):e0224899. doi:10.1371/JOURNAL.PONE.0224899

8. Dipanjan Sarkar. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. (Sarkar, ed.). apress; 2016.

9. Cammel SA, de Vos MS, van Soest D, *et al.* How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Med Inform Decis Mak*. 2020;20(1). doi:10.1186/S12911-020-1104-5

10. Minai AA, Doumit S, Minai A. Online News Media Bias Analysis using an LDA-NLP Approach. Published online 2012. Accessed September 13, 2022. https://www.researchgate.net/publication/267559458

11. Walsh J, Cave J, Griffiths F. Spontaneously Generated Online Patient Experience of Modafinil: A Qualitative and NLP Analysis. *Front Digit Health*. 2021;3. doi:10.3389/FDGTH.2021.598431

12. Greaves F, Laverty AA, Ramirez Cano D, *et al.* Tweets about hospital quality: A mixed methods study. *BMJ Qual Saf*. 2014;23(10):838-846. doi:10.1136/BMJQS-2014-002875

13. Hawkins JB, Brownstein JS, Tuli G, *et al.* Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*. 2016;25(6):404-413. doi:10.1136/BMJQS-2015-004309

14. Gonçalves M, Coheur L, Baptista J, Mineiro A. Evaluating computational resources for Portuguese. 12:51-68. doi:10.21814/lm.12.2.331

15. Diogo J, Ferreira C. Python para Pré-processamento e Extracção de Características a partir de Texto Português. Published online 2019.

16. CardioFollow.AI: Sistema inteligente para follow-up em cirurgia cardiotorácica - Value for Health CoLAB. Accessed October 18, 2022. https://vohcolab.org/pt/projects/cardiofollow-ai-sistema-inteligente-para-follow-up-na-cirurgia-cardiotoracica/

17. Bodulovic G, Wang S, de Morpurgo M, Saunders EJ. Telehealth around the world: A global guide | Insights | DLA Piper Global Law Firm. Accessed August 26, 2022. https://www.dlapiper.com/en/italy/insights/publications/2020/11/telehealth-around-the-world-global-guide/

18. Sara Carrasqueiro S (Portugal); AR, (Portugal); Ana Esteves S (Portugal); CP, (Portugal); Diogo Martins S (Portugal); LM, (Portugal); REPORT on EU state of play on telemedicine services and uptake recommendations Document Information.

19. Omboni S. Connected Health in Hypertension Management. *Front Cardiovasc Med*. 2019;6. doi:10.3389/FCVM.2019.00076

20. GHA Institute:Telemedicine and Telehealth 3 Day Online Certification Course - GHA Institute. Accessed August 26, 2022. https://www.ghainstitute.com/gha-institutetelemedicine-and-telehealth-3-day-online-certification-course/

21. Eren H (Professor of electrical engineering), Webster JG. *Telehealth and Mobile Health*. Accessed August 26, 2022. https://www.routledge.com/Telehealth-and-Mobile-Health/Eren-Webster/p/book/9781138893498

22. Bull C, Byrnes J, Hettiarachchi R, Downes M. A systematic review of the validity and reliability of patient-reported experience measures. *Health Serv Res*. 2019;54(5):1023-1035. doi:10.1111/1475-6773.13187

23. Lapin BR, Honomichl RD, Thompson NR, *et al*. Association Between Patient Experience With Patient-Reported Outcome Measurements and Overall Satisfaction With Care in Neurology. *Value Health*. 2019;22(5):555-563. doi:10.1016/J.JVAL.2019.02.007

24. Larson E, Sharma J, Bohren MA, Tunçalp Ö. When the patient is the expert: measuring patient experience and satisfaction with care. *Bull World Health Organ*. 2019;97(8):563-569. doi:10.2471/BLT.18.225201

25. Bevan N, Kirakowski J, Maissel J. What is Usability? Published online 1991.

26. Denzin NK, Lincoln YS. Introduction: The Discipline and Practice of Qualitative Research. - PsycNET. Accessed August 29, 2022. https://psycnet.apa.org/record/2005-07735-001

27. Mays N, Pope C. Introduction, The Role of Theory in Qualitative Research. In: Pope C, Bays N, eds. *Qualitative Research in Health Care*. 4th ed. John Wiley & Sons; 2020.

28. Mays N, Pope C. Quality in Qualitative Research. In: Mays N, Pope C, eds. *Qualitative Research in Health Care*. 4th ed. John Wiley & Sons Ltd; 2020:211-231.

29. Jurafsky D, James M. *Speech and Language Processing*. 3rd ed.; 2022.

30. Albalawi R, Yeap TH, Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front Artif Intell*. 2020;3:42. doi:10.3389/FRAI.2020.00042

31. Bird S, Klein E, Loper E. *Natural Language Processing with Python*. First. (Steele J, ed.). O'Reilly Media; 2009.

32. Wang X, Cao J, Liu Y, Gao S, Deng X. Text clustering based on the improved TFIDF by the iterative algorithm. *Proceedings - 2012 IEEE Symposium on Electrical and Electronics Engineering, EEESYM 2012*. Published online 2012:140-143. doi:10.1109/EEESYM.2012.6258608

33. Ghosh K, Banerjee A, Chatterjee S, Sen S. Imbalanced Twitter Sentiment Analysis using Minority Oversampling. In: *2019 IEEE 10th International Conference on Awareness Science and Technology (ICAST)*. IEEE; 2019:1-5. doi:10.1109/ICAwST.2019.8923218

34. Chawla N v, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357.

35. Younis Thanoun M, Yaseen MT. A Comparative Study of Parkinson Disease Diagnosis in Machine Learning. *ACM International Conference Proceeding Series*. Published online October 9, 2020:23-28. doi:10.1145/3441417.3441425

36. Gan J, Qi Y. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. *Entropy*. 2021;23(10). doi:10.3390/E23101301

37. Srividhya V, Shini G. (PDF) Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews. *Indian Journal of Natural Sciences*. Published online 2020. Accessed September 15, 2022. https://www.researchgate.net/publication/350236296_Comparison_of_LDA_and_NMF_Topic_Modeling_Techniques_for_Restaurant_Reviews

38. Aye YM, Liew S, Neo SX, *et al.* Patient-Centric Care for Parkinson's Disease: From Hospital to the Community. *Front Neurol*. 2020;11:502. doi:10.3389/FNEUR.2020.00502/BIBTEX

39. Davis K, Schoenbaum SC, Audet AM. A 2020 Vision of Patient-Centered Primary Care. *J Gen Intern Med*. 2005;20(10):953. doi:10.1111/J.1525-1497.2005.0178.X

40. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*. 2013;3(1):e001570. doi:10.1136/BMJOPEN-2012-001570

41. Tsianakas V, Maben J, Wiseman T, *et al.* Using patients experiences to identify priorities for quality improvement in breast cancer care: Patient narratives, surveys or both? *BMC Health Serv Res*. 2012;12(1):1-11. doi:10.1186/1472-6963-12-271/TABLES/3

42. Tyser AR, Abtahi AM, McFadden M, Presson AP. Evidence of non-response bias in the Press-Ganey patient satisfaction survey. *BMC Health Serv Res*. 2016;16(1). doi:10.1186/s12913-016-1595-z

43. Gualandi R, Masella C, Piredda M, Ercoli M, Tartaglini D. What does the patient have to say? Valuing the patient experience to improve the patient journey. *BMC Health Serv Res*. 2021;21(1):1-12. doi:10.1186/S12913-021-06341-3/TABLES/5

44. Gerteis M ELSDJDT. Through the Patient's Eyes: Understanding and Promoting Patient-Centered Care. *San Francisco, Calif: Jossey-Bass*. Published online 1993.

45. Kuipers SJ, Nieboer AP, Cramm JM. Easier said than done: Healthcare professionals' barriers to the provision of patient-centered primary care to patients with multimorbidity. *Int J Environ Res Public Health*. 2021;18(11). doi:10.3390/IJERPH18116057

46. Bridge E, Gotlib Conn L, Dhanju S, Singh S, Moody L. The patient experience of ambulatory cancer treatment: a descriptive study. *Curr Oncol*. 2019;26(4):e482-e493. doi:10.3747/CO.26.4191

47. Seleznev I, Alibekova R, Clementi A. Patient satisfaction in Kazakhstan: Looking through the prism of patient healthcare experience. *Patient Educ Couns*. 2020;103(11):2368-2372. doi:10.1016/J.PEC.2020.05.004

48. de Silva D. Measuring patient experience. *The Health Foundation Inspiring Improvement*. Published online 2013. Accessed August 30, 2022. https://www.health.org.uk/publications/measuring-patient-experience

49. Shaverdian N, Gillespie EF, Cha E, *et al.* Impact of Telemedicine on Patient Satisfaction and Perceptions of Care Quality in Radiation Oncology. *J Natl Compr Canc Netw*. 2021;19(10):1174-1180. doi:10.6004/JNCCN.2020.7687

50. Brown CM, Richards KM, Vohra Y, *et al.* Evaluation of access to care issues in patients with breast cancer. *J Med Econ*. 2021;24(1):38-45. doi:10.1080/13696998.2020.1858580

51. Okuda M, Yasuda A, Tsumoto S. An approach to exploring associations between hospital structural measures and patient satisfaction by distance-based analysis. *BMC Health Serv Res*. 2021;21(1). doi:10.1186/S12913-020-06050-3

52. Pacheco Barzallo D, Köhn S, Tobler S, Délitroz M, Gemperli A. Measuring patient satisfaction in acute care hospitals: nationwide monitoring in Switzerland. *Z Evid Fortbild Qual Gesundhwes*. 2021;165:27-34. doi:10.1016/J.ZEFQ.2021.07.001

53. Stahl K, Groene O. ASK ME!-Routine measurement of patient experience with patient safety in ambulatory care: A mixed-mode survey. *PLoS One*. 2021;16(12). doi:10.1371/JOURNAL.PONE.0259252

54. Jones CH, O'Neill S, McLean KA, Wigmore SJ, Harrison EM. Patient experience and overall satisfaction after emergency abdominal surgery. *BMC Surg*. 2017;17(1). doi:10.1186/S12893-017-0271-5

55. Tverdal CB, Howe EI, RØe C, *et al.* Traumatic brain injury: Patient experience and satisfaction with discharge from trauma hospital. *J Rehabil Med*. 2018;50(6):505-513. doi:10.2340/16501977-2332

56. Donaghy E, Atherton H, Hammersley V, *et al.* Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. *The British Journal of General Practice*. 2019;69(686):e586. doi:10.3399/BJGP19X704141

57. Poudel L, Baskota S, Mali P, *et al.* Patient Satisfaction in Out-patient Services at a Tertiary Care Center: A Descriptive Cross-sectional Study. *JNMA J Nepal Med Assoc*. 2020;58(225):301. doi:10.31729/JNMA.4917

58. Addington-Hall J. Survey research: methods of data collection, questionnaire design and piloting. In: *Research Methods e Palliative Care*. Oxford; 2005. Accessed August 29, 2022. https://books.google.pt/books?id=pf5QEAAAQBAJ&printsec=copyright&hl=pt-PT&source=gbs_pub_info_r#v=onepage&q&f=true

59. McGrath C, Palmgren PJ, Liljedahl M. Twelve tips for conducting qualitative research interviews. *Med Teach*. 2019;41(9):1002-1006. doi:10.1080/0142159X.2018.1497149

60. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. *Ann Fam Med*. 2017;15(3):225-229. doi:10.1370/afm.2095

61. Imlach F, McKinlay E, Middleton L, *et al.* Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. *BMC Fam Pract*. 2020;21(1):1-14. doi:10.1186/S12875-020-01336-1/TABLES/4

62. Chen Y, Chen Y, Zheng K, *et al.* A qualitative study on user acceptance of a home-based stroke telerehabilitation system. *Top Stroke Rehabil*. 2020;27(2):81. doi:10.1080/10749357.2019.1683792

63. Javanparast S, Roeger L, Kwok Y, Reed RL. The experience of Australian general practice patients at high risk of poor health outcomes with telehealth during the COVID-19 pandemic: a qualitative study. *BMC Fam Pract*. 2021;22(1). doi:10.1186/S12875-021-01408-W

64. Woo K, Dowding DW. Decision-making Factors Associated with Telehealth Adoption by Patients with Heart Failure at Home: A Qualitative Study. *Comput Inform Nurs*. 2020;38(4):204. doi:10.1097/CIN.0000000000000589

65. O'Cathain A, Drabble SJ, Foster A, *et al.* Being Human: A Qualitative Interview Study Exploring Why a Telehealth Intervention for Management of Chronic Conditions Had a Modest Effect. *J Med Internet Res*. 2016;18(6). doi:10.2196/JMIR.5879

66. Madden N, Emeruwa UN, Friedman AM, *et al.* Telehealth Uptake into Prenatal Care and Provider Attitudes during the COVID-19 Pandemic in New York City: A Quantitative and Qualitative Analysis. *Am J Perinatol*. 2020;37(10):1005-1014. doi:10.1055/S-0040-1712939

67. Kennedy NR, Steinberg A, Arnold RM, *et al.* Perspectives on Telephone and Video Communication in the Intensive Care Unit during COVID-19. *Ann Am Thorac Soc*. 2021;18(5):838-847. doi:10.1513/ANNALSATS.202006-729OC

68. Dejonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. *Fam Med Com Health*. 2019;7:57. doi:10.1136/fmch-2018-000057

69. Hinton L, Ryan S. Interviews. In: Pope C, Mays N, eds. *Qualitative Research in Health Care*. 4th ed. John Wiley and Sons Ltd.; 2020.

70. Sauers-Ford HS, Hamline MY, Gosdin MM, *et al.* Acceptability, Usability, and Effectiveness: A Qualitative Study Evaluating a Pediatric Telemedicine Program. *Acad Emerg Med*. 2019;26(9):1022-1033. doi:10.1111/ACEM.13763

71. Kitzinger J. Qualitative Research: Introducing focus groups. *BMJ*. 1995;311(7000):299-302. doi:10.1136/BMJ.311.7000.299

72. Tritter J, Landstad B. Focus groups. In: Pope C, Mays N, eds. *Qualitative Research in Health Care*. 4th ed. John Wiley & Sons Ltd.; 2020.

73. Portz JD, Bayliss EA, Bull S, *et al*. Using the Technology Acceptance Model to Explore User Experience, Intent to Use, and Use Behavior of a Patient Portal Among Older Adults With Multiple Chronic Conditions: Descriptive Qualitative Study. *J Med Internet Res*. 2019;21(4). doi:10.2196/11604

74. Busetto L, Wick W, Gumbinger C. How to use and assess qualitative research methods. *Neurological Research and Practice 2020 2:1*. 2020;2(1):1-10. doi:10.1186/S42466-020-00059-Z

75. Pope C, Ziebland S, Mays N. Analysis. In: Pope C, ed. *Qualitative Research in Healthcare*. 4th ed. John Wiley & Sons Ltd; 2020.

76. ATLAS.ti | The Qualitative Data Analysis & Research Software - ATLAS.ti. Accessed September 9, 2022. https://atlasti.com/pt

77. TextAnnotator: A web-based annotation suite for texts. Accessed September 9, 2022. https://dh2020.adho.org/wp-content/uploads/2020/07/547_TextAnnotatorAwebbasedannotationsuitefortexts.html

78. MAXQDA | All-In-One Tool for Qualitative Data Analysis & Mixed Methods - MAXQDA. Accessed September 9, 2022. https://www.maxqda.com/homepage-4

79. Ratanjee-Vanmali H, Swanepoel DW, Laplante-Lévesque A. Patient Uptake, Experience, and Satisfaction Using Web-Based and Face-to-Face Hearing Health Services: Process Evaluation Study. *J Med Internet Res*. 2020;22(3). doi:10.2196/15875

80. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101. doi:10.1191/1478088706QP063OA

81. Polinski JM, Barker T, Gagliano N, Sussman A, Brennan TA, Shrank WH. Patients' Satisfaction with and Preference for Telehealth Visits. *J Gen Intern Med*. 2016;31(3):269-275. doi:10.1007/S11606-015-3489-X

82. Yip MP, Chang AM, Chan J, Mackenzie AE. Development of the Telemedicine Satisfaction Questionnaire to evaluate patient satisfaction with telemedicine: a preliminary study. *J Telemed Telecare*. 2003;9(1):46-50. doi:10.1258/135763303321159693

83. Gibson-Helm M, Teede H, Dunaif A, Dokras A. Delayed Diagnosis and a Lack of Information Associated With Dissatisfaction in Women With Polycystic Ovary Syndrome. *J Clin Endocrinol Metab*. 2017;102(2):604-612. doi:10.1210/JC.2016-2963

84. Hentati F, Cabrera CI, D'Anza B, Rodriguez K. Patient satisfaction with telemedicine in rhinology during the COVID-19 pandemic. *Am J Otolaryngol*. 2021;42(3). doi:10.1016/J.AMJOTO.2021.102921

85. Ahmed F, Burt J, Roland M. Measuring patient experience: concepts and methods. *Patient*. 2014;7(3):235-241. doi:10.1007/S40271-014-0060-5

86. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth*. 2017;11(Suppl 1):S80. doi:10.4103/SJA.SJA_203_17

87. Comín Bertrán E. Validation of questionnaires. *Atencion primaria / Sociedad Española de Medicina de Familia y Comunitaria*. 1990;7(5):386-390. doi:10.1016/S2173-5743(09)70115-7

88. Rietz T, Maedche A. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research; Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. doi:10.1145/3411764.3445591

89. Wang T, Giunti G, Melles M, Goossens R. Digital Patient Experience: Umbrella Systematic Review. *J Med Internet Res 2022;24(8):e37952 https://www.jmir.org/2022/8/e37952*. 2022;24(8):e37952. doi:10.2196/37952

90. Mar Negreiro. The rise of digital health technologies during the pandemic. *European Parliamentary Research Service*. Published online 2021.

91. Monaghesh E, Hajizadeh A. The role of telehealth during COVID-19 outbreak: A systematic review based on current evidence. *BMC Public Health*. 2020;20(1):1-9. doi:10.1186/S12889-020-09301-4/TABLES/1

92. Knapp A, Harst L, Hager S, Schmitt J, Scheibe M. Use of Patient-Reported Outcome Measures and Patient-Reported Experience Measures Within Evaluation Studies of Telemedicine Applications: Systematic Review. *J Med Internet Res*. 2021;23(11). doi:10.2196/30042

93. Johnson BA, Lindgren BR, Blaes AH, *et al.* The New Normal? Patient Satisfaction and Usability of Telemedicine in Breast Cancer Care. *Ann Surg Oncol*. 2021;28(10):5668-5676. doi:10.1245/S10434-021-10448-6

94. Thomas NA, Drewry A, Racine Passmore S, Assad N, Hoppe KK. Patient perceptions, opinions and satisfaction of telehealth with remote blood pressure monitoring postpartum. *BMC Pregnancy Childbirth*. 2021;21(1). doi:10.1186/S12884-021-03632-9

95. Chang PJ, Jay GM, Kalpakjian C, Andrews C, Smith S. Patient and Provider-Reported Satisfaction of Cancer Rehabilitation Telemedicine Visits During the COVID-19 Pandemic. *PM R*. 2021;13(12):1362-1368. doi:10.1002/PMRJ.12552

96. Lechien JR, Radulesco T, Distinguin L, *et al.* Patient and otolaryngologist perceptions of telemedicine during COVID-19 pandemic. *Eur Arch Otorhinolaryngol*. 2021;278(10):4101-4105. doi:10.1007/S00405-021-06624-9

97. Liu CH, Goyal D, Mittal L, Erdei C. Patient Satisfaction with Virtual-Based Prenatal Care: Implications after the COVID-19 Pandemic. *Matern Child Health J*. 2021;25(11):1735-1743. doi:10.1007/S10995-021-03211-6

98. Sathiyaraj A, Lopez H, Surapaneni R. Patient satisfaction with telemedicine for prechemotherapy evaluation during the COVID-19 pandemic. *Future Oncol*. 2021;17(13):1593-1600. doi:10.2217/FON-2020-0855

99. Hajesmaeel-Gohari S, Bahaadinbeigy K. The most used questionnaires for evaluating telemedicine services. *BMC Med Inform Decis Mak*. 2021;21(1):1-11. doi:10.1186/S12911-021-01407-Y/FIGURES/2

100. Parmanto B, Lewis, Jr. AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). *Int J Telerehabil*. 2016;8(1):3-10. doi:10.5195/IJT.2016.6196

101. Hirani SP, Rixon L, Beynon M, *et al*. Quantifying beliefs regarding telehealth: Development of the Whole Systems Demonstrator Service User Technology Acceptability Questionnaire. *J Telemed Telecare*. 2017;23(4):460-469. doi:10.1177/1357633X16649531

102. Attkisson CC, Zwick R. The client satisfaction questionnaire: Psychometric properties and correlations with service utilization and psychotherapy outcome. *Eval Program Plann*. 1982;5(3):233-237. doi:10.1016/0149-7189(82)90074-X

103. Chin JP, Diehl VA, Norman KL. Development of an instrument measuring user satisfaction of the human-computer interface. *Conference on Human Factors in Computing Systems - Proceedings*. 1988;Part F130202:213-218. doi:10.1145/57167.57203

104. SUS: A "Quick and Dirty" Usability Scale. *Usability Evaluation In Industry*. Published online June 11, 1996:207-212. doi:10.1201/9781498710411-35

105. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. *Telemed J E Health*. 2009;15(9):830-839. doi:10.1089/TMJ.2009.0030

106. Bakken S, Grullon-Figueroa L, Izquierdo R, *et al*. Development, validation, and use of English and Spanish versions of the telemedicine satisfaction and usefulness questionnaire. *J Am Med Inform Assoc*. 2006;13(6):660-667. doi:10.1197/JAMIA.M2146

107. Thayaparan AJ, Mahdi E. The Patient Satisfaction Questionnaire Short Form (PSQ-18) as an adaptable, reliable, and validated tool for use in various settings. *Med Educ Online*. 2013;18(1). doi:10.3402/MEO.V18I0.21747

108. Finkelstein SM, MacMahon K, Lindgren BR, *et al*. Development of a remote monitoring satisfaction survey and its use in a clinical trial with lung transplant recipients. *J Telemed Telecare*. 2012;18(1):42. doi:10.1258/JTT.2011.110413

109. Demiris G. Principles of survey development for telemedicine applications. *http://dx.doi.org/101258/135763306776738549*. 2016;12(3):111-115. doi:10.1258/135763306776738549

110. Leonardsen ACL, Hardeland C, Helgesen AK, Grøndahl VA. Patient experiences with technology enabled care across healthcare settings- a systematic review. *BMC Health Serv Res*. 2020;20(1). doi:10.1186/S12913-020-05633-4

111. Ware JE, Snyder MK, Wright WR, Davies AR. Defining and measuring patient satisfaction with medical care. *Eval Program Plann*. 1983;6(3-4):247-263. doi:10.1016/0149-7189(83)90005-8

112. Alkire (née Nasr) L, O'Connor GE, Myrden S, Köcher S. Patient experience in the digital age: An investigation into the effect of generational cohorts. *Journal of Retailing and Consumer Services*. 2020;57:102221. doi:10.1016/J.JRETCONSER.2020.102221

113. Lewis JR. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *https://doi.org/101080/10447319509526110*. 2009;7(1):57-78. doi:10.1080/10447319509526110

114. López A, Detz A, Ratanawongsa N, Sarkar U. What patients say about their doctors online: a qualitative content analysis. *J Gen Intern Med*. 2012;27(6):685-692. doi:10.1007/S11606-011-1958-4

115. Doing-Harris K, Mowery DL, Daniels C, Chapman WW, Conway M. Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annual Symposium Proceedings*. 2016;2016:524. Accessed September 14, 2022. /pmc/articles/PMC5333198/

116. Bahja M, Lycett M. Identifying Patient Experience from Online Resources via Sentiment Analysis and Topic Modelling. Published online 2016. doi:10.1145/3006299.3006335

117. James TL, Villacis Calderon ED, Cook DF. Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Syst Appl*. 2017;71:479-492. doi:10.1016/j.eswa.2016.11.004

118. Floresta Sintática. Accessed October 27, 2022. https://www.linguateca.pt/Floresta/principal.html

119. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Qual Manag Health Care*. 2012;21(1):9-19. doi:10.1097/QMH.0B013E3182417FC4

120. spaCy · Industrial-strength Natural Language Processing in Python. Accessed September 20, 2022. https://spacy.io/?

121. NLTK :: Natural Language Toolkit. Accessed September 20, 2022. https://www.nltk.org/

122. Jiménez-Zafra SM, Martin-Valdivia MT. Analysis of patient satisfaction in Dutch and Spanish online reviews. Published online 2017. Accessed September 20, 2022. http://www.csn.cancer.org

123. Capelas M, Ferreira R. Diretório de escalas validadas para Português Europeu 2020. doi:http://doi.org/10.34632/9789725407837

124. User Experience Questionnaire (UEQ). Accessed October 6, 2022. https://www.ueq-online.org/

125. pandas - Python Data Analysis Library. Accessed October 27, 2022. https://pandas.pydata.org/

126. scikit-learn: machine learning in Python — scikit-learn 1.1.3 documentation. Accessed October 27, 2022. https://scikit-learn.org/stable/

127. Gensim: Topic modelling for humans. Accessed October 27, 2022. https://radimrehurek.com/gensim/

128. Matplotlib — Visualization with Python. Accessed October 27, 2022. https://matplotlib.org/

129. WordCloud for Python documentation — wordcloud 1.8.1 documentation. Accessed October 27, 2022. https://amueller.github.io/word_cloud/

130. pyLDAvis — pyLDAvis 2.1.2 documentation. Accessed October 27, 2022. https://pyldavis.readthedocs.io/en/latest/readme.html

131. NumPy. Accessed October 27, 2022. https://numpy.org/

132. RandomOverSampler — Version 0.9.1. Accessed October 10, 2022. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

133. Chen Y, Skiena S. Building sentiment lexicons for all major languages. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. 2014;2:383-389. doi:10.3115/V1/P14-2063

134. GitHub - doccano/doccano: Open source annotation tool for machine learning practitioners. Accessed October 10, 2022. https://github.com/doccano/doccano

135. Dey L, Chakraborty S, Biswas A, Bose B, Tiwari S. Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*. 2016;8(4):54-62. doi:10.5815/ijieeb.2016.04.07

136. Sarakit P, Theeramunkong T, Haruechaiyasak C. Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. *ICAICTA 2015 - 2015 International Conference on Advanced Informatics: Concepts, Theory and Applications*. Published online November 20, 2015. doi:10.1109/ICAICTA.2015.7335373

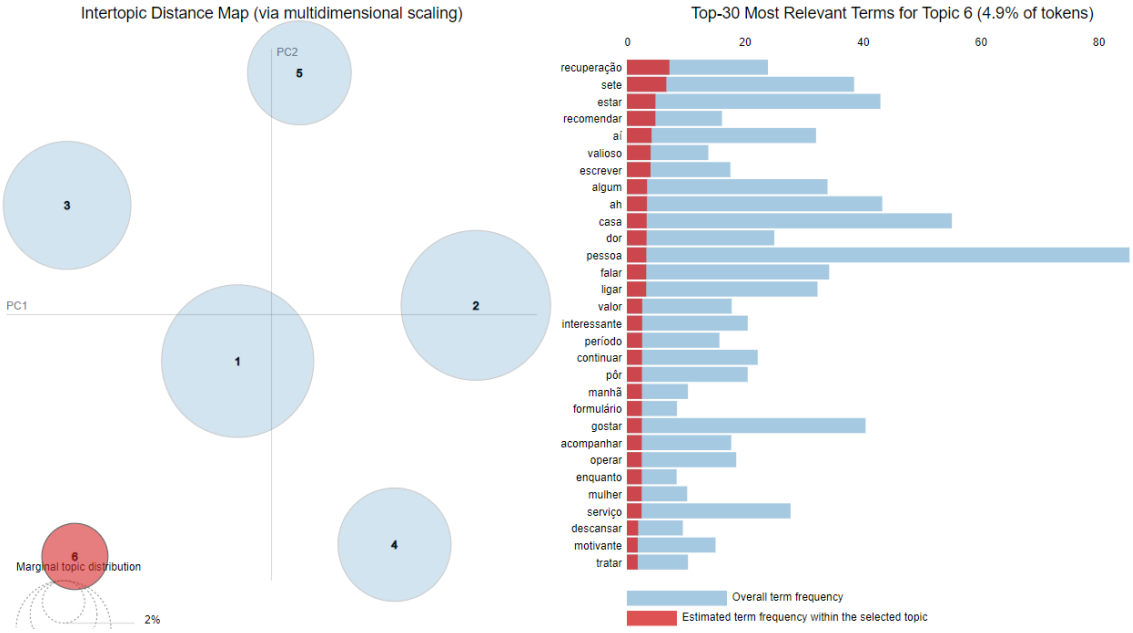
137. Jacobi C, van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*. 2016;4(1):89-106. doi:10.1080/21670811.2015.1093271

138. Allahyari M, Pouriyeh S, Assefi M, *et al*. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Published online July 10, 2017. Accessed October 18, 2022. https://www.researchgate.net/publication/338491281_Yoga-Veganism_Correlation_Mining_of_Twitter_Health_Data

139. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276. doi:10.11613/bm.2012.031

140. BERT Explained: State of the art language model for NLP | by Rani Horev | Towards Data Science. Accessed September 13, 2022. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

8 Appendix



Appendix A - LDA topic model vizualization using pyLDAvis. The selected topic model here is number 6. The frequency of each term within this topic is shown on the right.

Questionário de avaliação da experiência do doente em relação ao serviço digital

Este questionário foi desenvolvido pela equipa de investigação do laboratório colaborativo Value for Health CoLAB. Resultou da necessidade de termos um questionário, em língua portuguesa, que permitisse avaliar a experiência do doente em projetos de inovação de serviços digitais em saúde, com crescente relevância numa cultura de saúde baseada em valor. O seu desenvolvimento foi feito com base nas referências citadas no fim deste questionário e na nossa experiência na implementação de estudos em contextos reais de cuidados de saúde.

Pretendemos partilhar o questionário com equipas que possam contribuir para a sua validação, esperando receber comentários sobre a experiência da sua aplicação em diferentes contextos de cuidados de saúde.

O questionário poderá ser utilizado apenas para fins de investigação.

Assinale uma opção para cada uma das seguintes afirmações:

A. <u>Satisfação geral</u>

1. Senti-me bem acompanhado/a com este serviço.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

2. O serviço interfere com a minha rotina diária.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

3. O serviço deve ser recomendado a pessoas com uma condição de saúde semelhante à minha.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

4. Estou satisfeito com este serviço.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

B. <u>Satisfação com o equipamento</u>

5. A explicação dada no momento de entrega do equipamento foi suficientemente esclarecedora.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

6. Foi fácil aprender a utilizar o equipamento.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

7. O tempo que demoro a fazer as medições diárias é aceitável.

☐	☐	☐	☐	☐

Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

8. Sempre que me engano a usar o equipamento, consigo ultrapassar com facilidade.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

9. O equipamento é confiável e tem poucos problemas técnicos.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

10. A equipa é rápida a responder quando ocorrem problemas técnicos.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

11. O equipamento que recebi invade a minha privacidade.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

12. O equipamento faz-me sentir desconfortável física ou emocionalmente.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

13. O equipamento permite que os profissionais acompanhem melhor a minha condição à distância.

☐	☐	☐	☐	☐
Discordo totalmente	Discordo	Concordo	Concordo totalmente	Não sei

1. Bower P, Cartwright M, Hirani SP, *et al.* A comprehensive evaluation of the impact of telemonitoring in patients with long-term conditions and social care needs: Protocol for the whole systems demonstrator cluster randomised trial. *BMC Health Services Research*. 2011;11(1):1-12. doi:10.1186/1472-6963-11-184/FIGURES/7
2. Finkelstein SM, MacMahon K, Lindgren BR, *et al.* Development of a remote monitoring satisfaction survey and its use in a clinical trial with lung transplant recipients. *Journal of Telemedicine and Telecare*. 2012;18(1):42-46. doi:10.1258/jtt.2011.110413
3. Parmanto B, Lewis, Jr. AN, Graham KM, Bertolet MH. Development of the Telehealth Usability Questionnaire (TUQ). *International Journal of Telerehabilitation*. 2016;8(1):3-10. doi:10.5195/ijt.2016.6196

Investigador responsável: Ana Rita Londral (ana.londral@vohcolab.org).

Appendix B - Final version of the questionnaire to evaluate patient experience with digital healthcare services.