



# **Chronic Pain Assessment from Patient Reports**

**Diogo Afonso Pedro Nunes**

Thesis to obtain the Master of Science Degree in  
**Computer Science and Engineering**

Supervisors: Doctor David Manuel Martins de Matos  
Doctor Joana Maria de Pinho Ferreira Gomes

## **Examination Committee**

Chairperson: Doctor Francisco João Duarte Cordeiro Correia dos Santos  
Supervisor: Doctor David Manuel Martins de Matos  
Member of the Committee: Doctor Ana Rita Mendes Londral

**January 2021**



# Acknowledgements

This work is supported by my supervising team, at INESC-ID, Instituto Superior Técnico (IST), Universidade de Lisboa (ULisboa), and Faculdade de Medicina da Universidade do Porto (FMUP). The cooperation of Centro Hospitalar Universitário de São João (CHUSJ) was indispensable in the establishment of the network between health professionals and chronic pain patients, which resulted in the collection of the studied dataset, during the trying times of the COVID-19 pandemic. This work was also supported by numerous health professionals, scattered around the country, who made their time available to help further the attainment of insights about chronic pain manifestation and the linguistic expression.

I would like to personally thank my advisor, Professor David Matos, for all the time, patience, and dedication he poured into this work. A year and a half ago, we started an atypical journey in pursuit of the unknown, which would not have been possible without your guidance and belief in me. Your work ethic and dedication are truly an inspiration to all your students. It was the reason for my first approach, and it is the reason for my wish to keep working by your side.

I would also like to thank Professors Fani Neto and Joana Ferreira Gomes, my co-advisors, for accepting the seemingly impossible challenge with open arms, and dedicating their time to create and manage the fundamental network of health professionals and patients. It was not an easy job, especially in the context of a pandemic.

I am grateful for all of my family, for being my safety net and the wind in my back. You never once doubted me, and I am who I am because of you.

I want to thank my girlfriend, Catarina, for being there in all the moments and shaping my life in unimaginable ways. You pulled me up when everything seemed to be going downhill. You are the strongest of both of us.

Finally, but not least, I express all of my gratitude to my mother, who sets an example ever-inspiring to everyone in her life. For your unbelievable dedication, bravery, love to our family, and untiring motivation to conquer life itself, I thank you from the bottom of my heart. Mom, you will always be the reason for the spark in me.

Lisboa, January 27, 2021  
Diogo Nunes



To everyone that helped me  
through this journey.  
I know who you are.



# Resumo

A dor é uma experiência subjectiva e privada. Esta é influenciada pela matriz de percepção do sujeito, e só pode ser observada a partir do exterior através de expressões ou comportamentos de dor. O presente trabalho propõe o estudo da linguagem da dor como um tipo específico da sua expressão, modelando descrições de experiências de dor crónica a partir de entrevistas gravadas, transcritas, recolhidas num contexto de cuidados de saúde. Sob esta análise linguística, as descrições são agregadas pelos tópicos semânticos que cobrem, o que permite a caracterização dos tópicos semânticos tanto do paciente como da experiência dolorosa. A caracterização semântica é utilizada para prever parâmetros clínicos associados à manifestação da dor, especificamente, a patologia diagnosticada e a intensidade de dor.

Os resultados obtidos mostram que a incorporação de informação semântica externa, adquirida em colecções externas que não têm as limitações da nossa coleção, provou ser mais adequada do que as abordagens tradicionais de modelação de tópicos. Os resultados obtidos mostram também uma relação entre a linguagem da dor e a patologia diagnosticada, com uma precisão de  $\sim 80\%$ , numa metodologia de validação Leave-One-Out. Esta relação não foi encontrada ao prever a intensidade da dor auto-reportada. Foram identificadas e discutidas várias causas para esta observação, remontando à própria definição de percepção da dor.

Este trabalho é motivado pelo estudo do processo cognitivo que incorpora a experiência dolorosa, que determina que as dimensões emocional, psicossocial e sociocultural do sujeito com dor desempenham um papel específico na modulação da percepção da dor e do sofrimento e expressão correspondentes, e o estudo da linguagem da dor, que se mostra portadora de parte desta informação.





# Abstract

Pain is a subjective and private experience. It is influenced by the subject's perception matrix, and can only be observed from the outside through expressions or behaviors of pain. The present work proposes to study the language of pain as a specific type of expression, by modeling descriptions of chronic pain experiences from recorded, transcribed interviews, collected in a healthcare setting. Under this linguistic analysis, the descriptions are aggregated by the semantic topics they cover, which allows for the semantic topic characterization of both the patient and the painful experience. The semantic characterization is then used to predict clinical parameters associated with the manifestation of chronic pain, specifically, the diagnosed pathology and the self-reported intensity of pain.

The obtained results show that the incorporation of external semantic information, previously acquired in external collections that do not carry the limitations of ours, proved to be better adjusted than the traditional topic modeling approaches. The obtained results also show a relation between the language of pain and the diagnosed pathology, with an accuracy score of  $\sim 80\%$ . This relation was not found when predicting the self-reported intensity of pain.

This work is motivated by the study of the cognitive process that embeds the painful experience, which determines that the emotional, psychosocial, and sociocultural dimensions of the subject in pain play a specific part in modulating the perception of pain and corresponding suffering and expression, and the study of the language of pain, which is shown to carry part of this information.



# Palavras Chave Keywords

## *Palavras Chave*

Dor Crónica

Percepção de Dor

Avaliação Computacional de Dor

Modelação de Tópicos para Dor

Extração de Informação a Partir da Fala

## *Keywords*

Chronic Pain

Pain Perception

Computational Pain Assessment

Topic Models for Pain

Information Extraction from Speech



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Objectives . . . . .	1
1.2	Contributions . . . . .	2
1.3	Document layout . . . . .	2
<b>2</b>	<b>The Nature of Pain</b>	<b>5</b>
2.1	Cognitive aspects of pain . . . . .	7
2.1.1	Emotional state . . . . .	8
2.1.2	Beliefs and expectations . . . . .	8
2.1.3	Behavior . . . . .	8
2.1.4	Sociocultural context . . . . .	9
2.2	Expression of pain . . . . .	10
2.2.1	Linguistic expression and description . . . . .	10
2.2.2	Lexical profile . . . . .	11
2.2.3	Grammatical structure and semantics . . . . .	12
2.3	Summary . . . . .	13
<b>3</b>	<b>Pain and Language Analysis</b>	<b>15</b>
3.1	Text-based analysis . . . . .	16
3.1.1	Topic modeling . . . . .	16
3.1.2	Topic Modeling: Evaluation metrics . . . . .	19

3.1.3	Short-text topic modeling . . . . .	20
3.2	Audio-based analysis . . . . .	26
3.2.1	Frame and utterance level features . . . . .	27
3.2.2	General architecture and evaluation . . . . .	28
3.2.3	Speech emotion recognition . . . . .	29
3.3	Summary . . . . .	34
<b>4</b>	<b>Dataset Definition</b>	<b>35</b>
4.1	Collected information . . . . .	35
4.2	Data preparation . . . . .	39
4.3	Baseline dataset and challenges . . . . .	41
4.4	Summary . . . . .	42
<b>5</b>	<b>Characterization of the Population</b>	<b>43</b>
5.1	Topic modeling . . . . .	43
5.1.1	Text preprocessing . . . . .	44
5.1.2	Models . . . . .	45
5.1.3	Evaluation . . . . .	46
5.1.3.1	Interpretability metrics . . . . .	47
5.1.3.2	Clustering metrics . . . . .	47
5.2	Characterization . . . . .	49
5.2.1	Overall population . . . . .	50
5.2.2	Topic similarity clusters . . . . .	50
5.2.3	Demographic and clinical clusters . . . . .	51
5.3	Results and discussion . . . . .	51
5.3.1	Topic modeling . . . . .	52

5.3.1.1	Interpretability . . . . .	53
5.3.1.2	Clustering . . . . .	57
5.3.2	Characterization . . . . .	59
5.3.2.1	Overall population . . . . .	61
5.3.2.2	Topic similarity clusters . . . . .	65
5.3.2.3	Demographic and clinical clusters . . . . .	68
5.4	Summary . . . . .	74
<b>6</b>	<b>Prediction of Clinical Parameters</b>	<b>75</b>
6.1	Task definition . . . . .	75
6.1.1	Pathology classification . . . . .	75
6.1.2	Pain intensity classification . . . . .	76
6.2	Feature extraction . . . . .	76
6.3	Evaluation . . . . .	78
6.4	Results and discussion . . . . .	78
6.4.1	Pathology classification . . . . .	79
6.4.2	Pain intensity classification . . . . .	83
6.5	Summary . . . . .	85
<b>7</b>	<b>Conclusions and Future Work</b>	<b>87</b>
7.1	Conclusions . . . . .	87
7.1.1	Data collection . . . . .	87
7.1.2	Linguistic characterization . . . . .	88
7.1.3	Prediction of clinical parameters . . . . .	89
7.2	Future work . . . . .	90

<b>I</b>	<b>Appendices</b>	<b>99</b>
<b>A</b>	<b>Patient Profiling</b>	<b>101</b>
A.1	Population distribution . . . . .	101
A.1.1	Demographic distribution . . . . .	101
A.1.2	Clinical distribution . . . . .	102
A.1.3	Linguistic and paralinguistic analysis . . . . .	104
A.1.3.1	Inter-question . . . . .	105
A.1.3.2	Interviewer engagement . . . . .	106
A.1.3.3	Intra-question (by interviewer) . . . . .	106
A.1.3.4	Correlation with demographic and clinical features . . . . .	107
A.2	Profiling tool . . . . .	110
A.2.1	Linguistic domain . . . . .	110
A.2.2	Clinical domain . . . . .	111
A.2.3	Results . . . . .	112
<b>B</b>	<b>Figures and Tables</b>	<b>115</b>
B.1	Pathology classification . . . . .	115
B.2	Pain intensity classification . . . . .	116



# List of Figures

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Nature of Pain</b>	<b>5</b>
<b>3</b>	<b>Pain and Language</b>	<b>15</b>
3.1	Factor decomposition of the representational matrix . . . . .	18
3.2	LDA box diagram . . . . .	19
3.3	Attention-based weighted pooling NN . . . . .	31
3.4	Emotion recognition model architecture with domain fusion . . . . .	33
<b>4</b>	<b>Dataset Definition</b>	<b>35</b>
4.1	Data collection complementary form. . . . .	38
4.2	Data preparation pipeline . . . . .	39
4.3	Audio cluster assignments for speaker diarization . . . . .	40
<b>5</b>	<b>Characterization of the population</b>	<b>43</b>
5.1	Distribution of vocabulary probabilities. . . . .	46
5.2	Mean percentage of tokens in the vocabulary considered sufficiently similar. . . . .	52
5.3	PPMI score of each model, over a range of topics. . . . .	53
5.4	Topic model modularity. . . . .	55

5.5	Silhouette of each model, for 12 clusters. . . . .	57
5.6	Matrix $M$ for each topic model, and the corresponding sparsity score. . . . .	58
5.7	Mean topic mixture by question, in percentage. . . . .	63
5.8	Topic importance metrics of the whole population. . . . .	64
5.9	Topic co-occurrence in the whole population. . . . .	65
5.10	Projected documents clustering metrics for a varying number of clusters. . . . .	66
5.11	Projected documents on a 2D visualization with PCA and t-SNE. . . . .	67
5.12	Mean topic mixture of each cluster of patients. . . . .	69
5.13	Topic importance distribution by cluster. . . . .	70
5.14	Distribution of continuous demographic and clinical parameters on each cluster of patients. . . . .	71
5.15	Distribution of categorical parameters per cluster. . . . .	71
5.16	Mean mixture of topics by group of patients, per clinical and demographic pa- rameters. . . . .	73
<b>6</b>	<b>Predicting Clinical Parameters</b>	<b>75</b>
6.1	Mean accuracy score of each experiment in Table 6.6. . . . .	80
6.2	Accuracy score of each feature type in Table 6.3. . . . .	82
6.3	Accuracy score of each feature type in Table 6.3. . . . .	82
6.4	Mean accuracy score of all experiments in an ablative fashion. . . . .	84
A.1	Demographic parameters distribution over the age spectrum. . . . .	102
A.2	Distribution of the clinical parameters. . . . .	103
A.3	Distribution of the clinical parameters given age bins. . . . .	104
A.4	Distribution of the clinical parameters by interviewer. . . . .	105
A.5	Feature distributions across questions . . . . .	106

A.6	Feature distributions across interviewers . . . . .	107
A.7	Word count feature distribution by question and by interviewer. . . . .	108
A.8	Word rate feature distribution by question and by interviewer. . . . .	108
A.9	TF-IDF feature distribution by question and by interviewer. . . . .	109
A.10	Patient profile example components of the linguistic domain. . . . .	113
A.11	Example of a patient clinical panel. . . . .	113
A.12	Example of a complete patient profile. . . . .	114
B.1	Mean accuracy score of each experiment in Table 6.6 . . . . .	119
B.2	Accuracy score of each feature type in Table 6.3 . . . . .	120
B.3	Accuracy score of each feature type in Table 6.3 . . . . .	121
B.4	Mean accuracy score of all experiments in Table 6.6 in an ablative fashion. . . . .	122



# List of Tables

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Nature of Pain</b>	<b>5</b>
<b>3</b>	<b>Pain and Language</b>	<b>15</b>
<b>4</b>	<b>Dataset Definition</b>	<b>35</b>
4.1	Example of an annotation file with speaker segments identified and labeled. . . .	40
4.2	Dataset inputs and corresponding description. . . . .	41
<b>5</b>	<b>Characterization of the population</b>	<b>43</b>
5.1	Top 20 words more frequent, before and after preprocessing. . . . .	45
5.2	Considered topic models for the experimental results. . . . .	46
5.3	Top 5 words with higher cosine similarity scores for a random sample of tokens.	53
5.4	Extracted topics by each model with hand-assigned labels. . . . .	54
5.5	CluWords(FastText) with topic labels. . . . .	60
<b>6</b>	<b>Predicting Clinical Parameters</b>	<b>75</b>
6.1	Distribution of patients per pathology . . . . .	75
6.2	Distribution of patients per level of pain intensity. . . . .	76

6.3	Considered types of features to extract from a document collection. . . . .	77
6.4	Possible aggregations that represent each patient by a single vector. . . . .	77
6.5	Text parameters of the experiments. . . . .	78
6.6	Configuration of all experiments. . . . .	78
A.1	Number of subjects per demographic parameter. . . . .	101
A.2	Number of subjects per identified pathology. . . . .	102
A.3	Correlation between select clinical and demographic parameters. . . . .	103
A.4	Correlation between clinical and demographic parameters with linguistic features. . . . .	107
B.1	Mean accuracy score of each experiment in Table 6.6. . . . .	115
B.2	Accuracy score of each feature type in Table 6.3. . . . .	115
B.3	Accuracy score of each feature type in Table 6.3. . . . .	116
B.4	Mean accuracy score of each experiment in Table 6.6. . . . .	116
B.5	Accuracy score of each feature type in Table 6.3. . . . .	117
B.6	Accuracy score of each feature type in Table 6.3 . . . . .	118

# 1 Introduction

Pain is a subjective and private experience. It is subjective because it is dependent on biomedical, psychological, and sociocultural dimensions, encompassing the perception matrix of the patient, that directly influence how it is perceived and consequently expressed by the subject in pain. Pain is also private, because if it is not expressed to the outside world, it cannot be observed and assessed. In this sense, the expressions of pain function as a window, allowing external entities to interpret and evaluate an otherwise private experience. Expressions of pain range from facial expressions, to verbal descriptions and changes in behavior. These, together with demographic and clinical parameters related to the manifestation of pain, are the inputs used by health professionals to assess and manage these patients.

Pain assessment and management are, arguably, complex tasks. Not only are they dependent on verbal and non-verbal communication established with the subject in pain, but also on the interpretation of this communication performed by the health professional. After years of experience, health professionals are capable of developing a model of pain, by learning how to associate certain key expressions to underlying states. Computationally analyzing expressions of pain may provide insights about the intrinsic characteristics of the experience, to ultimately aid health professionals with better pain management procedures.

## *1.1 Objectives*

An experience of pain is dependent on the perception matrix of the subject experiencing it. Language of pain, a specific type of expression, conveys information both about the perception of the subject and the underlying pain mechanisms, which are relevant details for an adequate pain management. Thus, the analysis of the language of pain, specifically through verbal descriptions of the experience in a healthcare context, may help develop a computational linguistic and paralinguistic model of pain, which in turn can be used to evaluate those descriptions and the dimensions of pain.

The hypothesis for this approach, from a linguistic standpoint, is that semantically related descriptions of pain may represent related experiences and can indirectly characterize the different types of pain. The paralinguistic perspective is also taken into account in order to capture the relevant communicative information not passed through text.

Concretely, the objective of the present work is two-fold, given a collection of descriptions of pain. First, to obtain a characterization of the population in the linguistic domain, and, second, to use said domain to predict clinical parameters related to the manifestation of pain.

## 1.2 Contributions

The present work contributed to the definition and implementation of a data collection protocol, as well as the preparation and definition of a baseline dataset of verbal descriptions of pain in Portuguese. To the extent of our knowledge, there is no previous dataset that merged verbal descriptions of pain in a healthcare context with demographic and clinical parameters, including the intensity of pain parallel to the description. This was possible with the establishment of a network between health professionals and patients suffering of chronic pain. The development of this work also lead to the submission of a financed project to Fundação para a Ciência e Tecnologia (FCT). Finally, 3 papers resulting from this work are under development, respectively focusing on the linguistic characterization of verbal descriptions of pain based on topic modeling, the usage of linguistic features for pathology prediction, and the merging of linguistic and clinical parameters to aid the management of pain in a healthcare context.

## 1.3 Document layout

The document is structured as follows.

Chapter 2 discusses the nature of pain, presenting the types of painful stimuli and characterizing the experiences of pain, as well as an in-depth look at the cognitive process involved in perceiving and expressing pain to the outside world, specifically examining the language of pain, the tool used to construct the descriptions of pain under study.

Chapter 3 briefly studies the methods and instruments used for a medical assessment of pain, and presents a discussion of the state-of-art of the corresponding computational linguistic



and paralinguistic methods.

Chapter 4 defines the dataset used in this work. It encompasses both the data collection protocol as well as the preparation pipeline, which produces the baseline dataset for the performed experiments. The challenges associated with the nature of the data are also discussed.

Chapters 5 and 6 present the experimental setup, results, and corresponding discussion of the main objectives, respectively, the characterization of the population on the linguistic domain, and the usage of said characterization to predict demographic and clinical parameters.

Finally, Chapter 7 presents final considerations regarding the whole work, and finalizes the study with future work.

This work is followed by two appendices. In Appendix A is made an exposition of the distribution of the population according to select demographic and clinical parameters, as well as linguistic and paralinguistic features. This is presented to help contextualize the core of the work. This report is followed by the proposition of a tool to automatically produce a patient profile which merges the clinical and linguistic domains. Finally, in Appendix B are tabulated the actual values of figures used in the discussion, as well as other figures deemed relevant for presentation, but not for discussion.



# The Nature of Pain

Pain is a sensation and an experience that issues a warning that something is probably wrong with the body. This sensation is exclusively private. The experience of pain resulting from that sensation is molded by a set of multi-domain factors, both individual and sociocultural. This experience is effectively the result of a complex cognitive process which takes as input noxious signals, the sense of self and the psychological, behavioral, and sociocultural embeddedness of the subject in pain. The cognitive process of pain can therefore be separated into two major components, the noxious signal and the subjective resulting experience.

The noxious signal, or painful stimulus, can be broadly classified into two categories, the physiological and the pathological. The physiological category encompasses both the nociceptive and inflammatory pains which are associated with sensory input from potential or actual tissue damage, respectively. Their purpose is two-fold: firstly to alert and protect the body from potential tissue damage, resulting in non-controlled bodily actions and reflexes, and, secondly, to discourage contact and movement involving the damaged tissue, serving the purpose of assisting in the healing process. On the other hand, the pathological category encompasses both the dysfunctional and neuropathic pains, which do not serve a specific function for well-being and survival and are presumably the result of maladaptation. This category of pain is commonly identified as a disease of the nervous system, amplifying, or generating sensory signals that should not be there ([Woolf, 2010](#)).

The experience of pain is triggered in a range of physiologically, psychologically, and emotionally unbalanced states, depending on the noxious stimulus, its temporal pattern of activity, and other factors. This is further influenced by the patient's perception of the pain, and consequent suffering and behavior.

An acute pain experience is usually associated with tissue damage, inflammation, and brief disease processes, thus encompassing both the nociceptive and inflammatory types of pain. The subject suffering this pain understands it to be essential for survival, functioning as

a warning that something is not right (Fink, 2000). The healing processes eventually overcome the injury and pain generally disappears with the elimination of the causal agent and inflammation (Dias, 2007). This usually takes a few days or weeks. Pain that persists for months or years is not classified as an acute pain experience (Loeser & Melzack, 1999). Due to its well-defined characteristics and short period of activity, assessment of acute pain can be a straightforward clinical practice (Breivik et al., 2008), and, thus, is not considered in the present study.

A chronic pain experience, in contrast, is characterized by its persistent state, either continuous or recurring, lasting for months, years, or a lifetime. The organism arrives at this state when the original damage overwhelms the healing processes, preventing the nervous system from restoring itself to the original state (Loeser & Melzack, 1999). Taking the perspective of pathological pain, it is commonly associated with a disease process, such as arthritis, cancer, and fibromyalgia (Fink, 2000), and can be perpetuated and intensified by factors other than the causal agent, such as stress, environment, culture, and affection (Loeser & Melzack, 1999). This experience can be expressed in a multitude of ways which are consequently dependent on the cultural, behavioral, and psychosocial dimensions of the subject in pain (Dansie & Turk, 2013), rendering it impossible to impartially experience, describe, and interpret pain as a pure noxious stimulus that would directly point to the causal agent and facilitate its mitigation. Assessment of persistent pain is, therefore, a demanding task, and considering that sometimes there is no identifiable objective pathology, most of the time it can only be based on the patient's explicit communication, both verbal and nonverbal. This process requires a comprehensive set of methodologies besides the standard pain assessment techniques, including a complete review of the patient's history and medical examination, and a set of screening and psychological interviews (Dansie & Turk, 2013) to effectively characterize all dimensions of the pain experience. Despite advances in research, chronic pain assessment and consequent management are still challenging (Loeser & Melzack, 1999; Fink, 2000; Breivik et al., 2008; Azevedo, Costa-Pereira, Mendonça, Dias, & Castro-Lopes, 2012).

Chronic pain is recognized as a major health problem, with impacts not only on the individual, but also on the social and economical levels. In 2012, it was concluded that 37% of the adult Portuguese population suffered of chronic pain (Azevedo et al., 2012), which was further estimated (at 31%) to have 738.85 million euros in associated costs, both directly and indirectly (Gouveia & Augusto, 2011). On the individual level, chronic pain dominates a multi-

tude of aspects of the patient's life, most of the time even extending to family and friends. This experience is usually accompanied by a chain of alterations on both the somatic and psychological levels. For instance, due to immobility, muscles start to weaken, discouraging further exercise or movement. This cycle can lead to sleep disturbances and a vulnerable immune system, effectively affecting the subject's psychological balance, leading to medication dependency, disability, isolation, depression and, sometimes, suicide (Dias, 2007). Adequate chronic pain assessment determines the quality of its management, which has been identified as a key procedure in improving the quality of life of these patients (Fink, 2000).

## 2.1 *Cognitive aspects of pain*

How the painful experience is perceived and conceptualized directly influences how it is expressed and consequently evaluated by an external entity (Dansie & Turk, 2013), which demands a comprehensive assessment of the patient as a whole. Therefore, the cognitive process of pain must be defined so that it may be possible to identify which factors influence this perception and corresponding suffering, and understand how it is expressed to the outside world.

The International Association for the Study of Pain (IASP) defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (Merskey & Bogduk, 1994). This definition relates the sensory input with the omnipresent experience. The relational element is the neuromatrix, which was defined by Melzack (2001) as "a widespread network of neurons that generates patterns, processes information that flows through it and ultimately produces the pattern that is felt as a sense of self". This modulating network encompasses past experiences, memories, and other factors such as culture and psychosocial states, outputting the multiple dimensions of the experience of pain together with regions of the brain involved in affective and cognitive activities (Loeser & Melzack, 1999).<sup>3</sup> In essence, sensory inputs are fed into the neuromatrix which then generates the perception and experience of pain based on the sense of self of the subject, adding a subjectivity filter to the experience. As stated before, the perception of pain is determined by a set of intrinsic personal factors, which range from past experiences and memories to emotional, psychological, sociocultural, and behavioral contexts. Determining each of these values for the patient in question will help characterize the private experience and underlying mechanisms of that pain.

### 2.1.1 Emotional state

The emotional state has been observed to influence the feeling and perception of pain, particularly in patients whose painful signal generates an imbalanced emotional state, which leads to negative responses to the experience of pain, causing more suffering, effectively perpetuating this cycle (Hansen & Streltzer, 2005). Specifically, the distress factors that are known to commonly worsen the pain experience are depression and anxiety. Depressive symptoms are known to intensify the perceived feeling of pain: a Portuguese population-based study determined that a lifetime history of these symptoms is significantly associated with chronic pain (Azevedo et al., 2012). Anxiety caused by attentional disorders (hypochondria) can lead subjects who are over-vigilant about bodily sensations to amplify them to the point of actually feeling painful, even when there was no noxious stimulus in the first place (Hansen & Streltzer, 2005). As a consequence, this may lead to fear and further disability. On the other hand, the opposite is expected when the subject is in a more positive emotional state or less focused on the debilitating factor of pain.

### 2.1.2 Beliefs and expectations

The way the patient feels and thinks about the disease, pain, and treatment processes psychologically modulates the experience, leading to commitment and empowerment issues towards the pain. Negative expectations and wrong beliefs about the recovery processes and the disease/pain itself may lead to an inaccurate perception matrix. Specifically, negative social cues, such as messages that communicate lack of confidence (especially on the side of the health professional), the act of prescribing medication (Hansen & Streltzer, 2005), or the sharing of false information in social circles, can all negatively impact the experience. Conversely, positive expectations and adequate beliefs may lead to a better engagement of the patient with the recovery processes and helping in overcoming the debilitating factors of pain.

### 2.1.3 Behavior

Factors such as the patient's disability, coping efforts, and communication are directly linked with the development of the experience. Particularly, it has been found that patients who were less satisfied with their pain management had significantly higher disability (Azevedo et al.,

2012). Hansen and Streltzer (2005) conclude that a crucial step for a better pain management is to convince the patient that an active role must be taken in retaking control of life (empowerment) and, consequently, minimizing the influence that pain has in the quality of life. However, this will be determined by the patient's openness about the whole experience and exposition of anxiety, beliefs, or doubts about the disease and treatments.

#### 2.1.4 Sociocultural context

On a larger social scale, there are some sociocultural parameters which have been suggested to influence the experience of pain in some way, namely, religion, ethnicity, and nationality (Miyahara, 2019). On the other hand, in a more fine-grained scale, it is expected that in specific social contexts expressions of painful experiences might vary, for instance, between friends, where the goal might be to open up about personal problems, or in a healthcare setting, where the primary goal is to seek medical help in mitigating the causal agent (Ehlich, 1985). Particularly, the enactive approach proposes pain as an embodied experience in a particular environment, which does not only affect its perception but the very experience itself (Miyahara, 2019). Population-based studies have also been able to identify correlations between demographic and social variables with chronic pain and associated disability (Azevedo et al., 2012). Regarding the education variable, a higher level is found to be correlated with an emphasized positive thinking (Leino-Kilpi, Maenpaa, & Katajisto, 1999), which in turn can positively affect the psychosocial and behavioral dimensions of the subject and experience. Conversely, weak foundational studies can lead to incoherent reflections about the disease, the spread of false information, and general misunderstanding of the complex processes, which result in the problems exposed in previous sections. In terms of employment, retired and unemployed subjects have been found to be statistically more susceptible to chronic pain. An immediate cause can be attributed to the lack of entertainment and diverse social contexts, leading to anxiety, fear, and inadequate beliefs. Finally, on the demographic axis, regarding age, it has been evidenced that older subjects are more likely to be associated with chronic pain and disability. Azevedo et al. (2012) have also found chronic pain and disability to be more associated with the feminine gender, than the masculine gender.

## 2.2 *Expression of pain*

A painful experience is private to the subject in pain and ultimately non-existent in an external entity's eyes. This experience is only accessible to the outside world through an outward expression or behavior, rendering the expression a necessary part of pain. For an external entity, by observing these expressions, it may be possible to infer the existence of pain in a quantified manner (Loeser & Melzack, 1999). Given that pain is a socioculturally embedded experience, and as a multitude of experiences and memories are accumulated, these expressions are eventually associated with specific types and intensities of pain. Furthermore, it is learned which behaviors are adequate for a given social context, from positive and negative reinforcement, which are the ones that produce the (seemingly) best outcome for a given painful experience (Hansen & Streltzer, 2005), and, ultimately, a context-dependent pain-to-expression transformation function is developed, which is inversely used to interpret someone's pain behavior.

The most common expressions of pain are cries, facial expressions, verbal interjections, descriptions, emotional distress, disability, and other behaviors that come as a consequence of these, such as lack of social interaction, exercise, movement, and productivity. The expression that is the object of study of the present work is the verbal description of the experience of pain, which includes both linguistic and paralinguistic aspects. The description oftentimes includes valuable information about the bodily distribution of the feeling of pain, temporal pattern of activity, and intensity. Additionally, the choice of words may reflect the underlying mechanisms of the causal agent (Wilson, Williams, & Butler, 2009), which in turn can be used to redirect the therapeutic processes. The language of pain is the tool used to build this description. Understanding this tool and how it is used for specific types of experiences allows us to build a linguistic and paralinguistic model of pain descriptions.

### 2.2.1 **Linguistic expression and description**

In a healthcare context, in order to get the desired help, the subject in pain must outwardly express and describe the underlying pain in the way that is felt to best expose the relevant factors of the experience. An expression is the action of disclosing one's current thoughts or feelings. Therefore, studying the expression of pain allows for the evaluation of the present sensation and experience. The different forms of expression that are relevant in a verbalized experience



are the emotional distress, the affection towards what is being said (for instance, when talking about the therapeutic processes), and interjections. In general, these manifestations are unintentional. On the other hand, one can (intentionally) describe any given interval of time of an experience of pain, past or present. However, as it has been characterized before, pain is an embodied response to a certain situation (Miyahara, 2019), rendering it simultaneously dependent on the subject and the context in which it is experienced. Thus, when it is to be described afterwards, both the subject and the context will most likely have changed, which can, in turn, result in inaccurate perceptions and descriptions. On top of this, the cognitive process of building these descriptions is limited both by the language (the tool) and the linguistic capability of the speaker to use that tool (Ehlich, 1985).

### 2.2.2 Lexical profile

In the realm of pain descriptions, there are specific words or combinations of words that have been found to be consistently associated with certain pains and intensities across subjects of different backgrounds.

The work initiated by Melzack and Torgerson (1971) aggregated these words, denominated pain descriptors, and performed a series of studies in order to categorize and relate them with pain indices which would be valuable for pain assessment. The main challenge was natural language ambiguity, where, for instance, many of these descriptors could be interpreted as synonyms with varying intensity, while others were only subtly different. These slight differences had to be captured and quantified, as they might represent specific characteristics of the pain the patient was trying to verbalize. Accordingly, pain descriptors were mainly categorized into three classes: sensory descriptors describe the sensation of the pain, for instance, burning, throbbing, stabbing; affective descriptors represent the emotional affection towards the pain, for instance, sickening, suffocating; and, finally, evaluative descriptors, provide a personal evaluation of the overall experience, for instance, mild, annoying, unbearable. For each class, the descriptors were organized and put on an intensity scale so that they could be compared and quantified. The result of these studies was the McGill Pain Questionnaire (MPQ), which is nowadays widely used to characterize pain from a verbal standpoint, having been demonstrated to provide reliable, valid indices of pain in a relatively efficient way (Katz & Melzack, 1992).

However, the identified MPQ descriptors represent only a portion of the lexical profile of the language of pain. A trivial analysis reveals the validity of this statement as translation between different languages is not always a one-to-one process, which means that different languages may use different words (or combinations of words) to describe the exact same characteristic. Accordingly, the MPQ has been adapted to different languages based on population-based studies (Stein & Mendl, 1988; Pimenta & Teixeira, 1996). On a second-order analysis, it is understood that language, culture, and social context are intertwined, influencing one another. Therefore, the possibility exists that in the same language (for example, Portuguese) the same characteristic of pain may be described by different words by patients of different sociocultural backgrounds. Some studies have specifically stated that the fixed quality of the MPQ ultimately limits the assessment in terms of stability and predictiveness, concluding that the descriptors should be subordinated to that sociocultural, linguistic background (Sullivan, 1995).

The study of the lexical profile of the language of pain suggests that there are language-specific pain descriptors which bear crucial information regarding the qualities of the underlying pain, that can be compared and quantified to output a pain index. It suggests that the patient's choice of words might be contributing to modulating the experience of pain and triggering cyclic worsening experiences (Wilson et al., 2009), and that the vocabulary is in fact an open set that can change over time and be different in certain sociocultural contexts. Thus, it is concluded that pain assessment from a verbal perspective would greatly benefit from an evaluative analysis that is flexible to the descriptors that the subject in pain feels that more adequately describe that unique pain experience.

### 2.2.3 Grammatical structure and semantics

Halliday (1998) performed a functional account of English pain descriptions in order to provide a theoretical background for the construction of pain descriptions: the Grammar of Pain. Halliday used the grammatical system of transitivity as the basis of his work. This system states that the world of experience is construed into a manageable set of process types which describe reality. In Halliday's setup, a process integrates three components: the process itself, which is realized by the verbal phrase; the participant, which is realized by the nominal phrase; and, finally, the quality, which may be realized by the adjective, adverbial, or prepositional phrases. With this framework at hand and given descriptions from patients, Halliday found the concept

of pain to be realized by any of the elements of a process, for instance (examples adapted from [Sussex \(2009\)](#)), “The wound is painful” (qualitative adjective), “I feel a lot of pain” (participant), and “My knee is hurting” (verb).

Based on Halliday’s framework, [Lascaratou \(2007\)](#) performed a corpus-based study composed of 69,996 words from 131 different patients naturally discussing their pain. She found the concept of pain realization as a verb in the process to be the most common, specifically arguing that this is the linguistic structure most associated with intense, personal expression/description of pain ([Sussex, 2009](#)).

## 2.3 Summary


In this chapter the nature of pain was briefly presented, identifying the experience as the result of the interpretation and expression of a given noxious signal, embedded in the subject’s emotional, psychological, and sociocultural dimensions, as well as the context in which pain is experienced. Both acute and chronic pain experiences were detailed, with focus on the challenges associated with the assessment and management of persistent pain, continuous or recurring.

Then, we dived into the cognitive aspects of the experience of pain, with the primary goal of identifying the key factors that have been shown to modulate the experience of pain, as well as to understand how they take influence. Specifically, we looked at the emotional state, beliefs and expectations, behavior, and sociocultural context as primary cognitive aspects. All of these factors are commonly identified by health professionals in order to assess the patient as whole, and provide a more adequate management of their pain.

Finally, we explored the language of pain as a specific type of expression of pain, and how the usage of this tool conveys relevant information about the patient and the underlying mechanisms of pain, as well as how it modulates the actual experience. We identified how certain types of words and expressions are associated with specific qualities and intensities of pain, and we also explored how the grammatical construction of a description of pain relates to the embodiment of the experience.



# Pain and Language Analysis



Pain assessment is the cornerstone of its management. An adequate assessment will provide significant insights to the extent and magnitude of the disease, and the development of the recovery process. From a medical standpoint, a generic approach to evaluation tries to understand the pattern of activity, intensity, and bodily distribution of the pain. A complete pain history includes these factors, additional descriptive qualities, and physical examinations, even though there is no direct linear relationship between the amount of detectable physical pathology and the reported pain intensity. This documentation is often helpful in revealing important aspects of co-morbidities ([Breivik et al., 2008](#)), in assessing the necessity of additional tests, and in providing a safeguard against over-interpretation of other findings, effectively characterizing the biomedical dimension of pain ([Dansie & Turk, 2013](#)). Widely accepted tools for an adequate collection of these data include the previously discussed MPQ ([Melzack, 1975](#)), the Brief Pain Inventory (BPI) ([Cleeland & Ryan, 1994](#)), and a number of other questionnaires and scales, such as the Visual Analogue Scale (VAS), and the Numeric Rating Scale (NRS) of pain intensity. In order to account for all the remaining dimensions of the experience, chronic pain assessment must also be accompanied with psychological and behavioral examinations. A screening interview (and a set of other more specialized interviews, if deemed necessary) is performed so that the health professional may be able to understand the patient's subjective filter through which pain is being perceived, felt, and expressed. This includes studying the patient's behavior alone and with significant others, assessment of the emotional state, explicit beliefs, and expectations ([Dansie & Turk, 2013](#)).

A linguistic analysis of the description of pain may provide insights on the aforementioned relevant factors to the assessment. Specifically, similar descriptions of pain might describe similar characteristics of different experiences of pain. Allowing these descriptions to be characterized by their semantic topics allows to quantify the relations between different experiences in this abstract space of semantic concepts, determining how similar they are. Additionally, it may be possible to characterize specific types of pain by their associated semantic topics.

A paralinguistic analysis, on the other hand, provides methods to evaluate the patient's verbalization of pain without necessarily considering the actual semantic or structural contents of the description. Specifically, by identifying the speaker's emotional state it may be possible to quantify the level of perception distortion that it may be causing and provide the health professional with a "second opinion". The following sections present the state-of-art regarding both types of analysis.

### 3.1 *Text-based analysis*

The analysis of syntactic and semantic structures of textual descriptions of pain may yield correlations between the content of the descriptions and other relevant medical or non-medical aspects of the painful experience. This includes the identification of the most significant descriptors or qualifying attributes, the aggregation of descriptions focusing on the same or similar concepts, sentiment analysis, and regression of any value from a description. This analysis may be performed with a multitude of methods and models. Specifically, topic models are capable of extracting semantic information from text in an unsupervised manner without relying on the explicit analysis of syntactic structures. The latter characteristic is especially relevant in contexts such as transcriptions of natural speech, which, in general, include repetitions, corrections, and other syntactically disruptive speech disfluencies not commonly present in written text. Thus, the text-based analysis of descriptions of pain, which inherit the aforementioned syntactically disruptive artifacts, will focus on topic modeling.

#### 3.1.1 **Topic modeling**

This task focuses on extracting implicit (latent) information in a given document from a collection, explicitly representing it with that information. Thus, each document is projected into the latent space of (abstract) semantic concepts of the collection, where the value of each dimension represents the weight of that latent topic in the given document. A topic is a cluster of weighted words, where the weight indicates the level of relevance that word has in the topic in such a way that the top relevant words of a topic are syntactically and/or semantically related, given the collection. Pragmatically, topic modeling can be seen as a dimensionality reduction technique as it provides a representation of documents in the lower-dimensionality space of latent

topics, which is usually much smaller than the vocabulary space. By itself, this task provides a new perspective on the documents and the collection, allowing for new measures of similarity, composition, and aggregation. This can then be used to enhance other tasks dependent on document representation, such as document classification, indexing, and clustering. Additionally, topics can be characterized by themselves when they are attributed with "meaning", given the context of a problem. In the following exposition, the following concepts will be used: the vocabulary  $V$ , of size  $|V|$ , is the set of words of a document collection, where each term (or word) is denoted  $w$ ; a document is a sequence of  $N$  terms, denoted  $\mathbf{w} = (w_1 w_2 \dots w_N)$ ; and a collection of  $M$  documents is denoted  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ .

Topic modeling methods follow either probabilistic or non-probabilistic approaches. Non-probabilistic approaches follow three steps, specifically, data representation, latent topic decomposition, and topic extraction. Common document collection representations are the term-document term frequency matrix  $N_{|V| \times M}$  and the Term Frequency Inverse Document Frequency TFIDF $_{|V| \times M}$  matrix. The former defines each entry  $n_{ij}$  as the number of times the vocabulary term  $w_i$  appears in document  $\mathbf{w}_j$ , and the latter each entry  $\text{tfidf}(t, d)$  as the frequency of the term  $t$  in document  $d$ , as previously defined by entry  $n_{ij}$ , also denominated  $\text{tf}(t, d)$ , multiplied by the inverse document frequency of that term in the whole collection, as defined in Eq. (3.1).

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \log \frac{M}{|\{d \in D : t \in d\}|} \quad (3.1)$$

The data representation matrix is then decomposed into low-rank ( $k$ ) factor matrices and, finally, the topic extraction step, which is dependent on the data representation matrix and consequent decomposition, is performed. For instance, the non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999) model performs a factorization of the form  $N \approx WH$  as in Figure 3.1a, constraining these factors to be non-negative, which encodes the intuition that a document is an additive combination of topics rather than complex cancellations between positive and negative factors, which would result from applying the (similar) dimensionality-reduction techniques of Principal Component Analysis (PCA) (Wold, Esbensen, & Geladi, 1987) or Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Thus, for this model, there is not the necessity for an explicit topic extraction step since the resulting matrices  $W$  and  $H$  represent, respectively, terms in the latent  $k$ -topic space and the

distributions of topics per document. Matrix  $W$  is sparse, intuitively representing the notion of a semantic topic and dictated by the resulting matrix  $H$ . Document variability is obtained by combining these parts/topics into a whole. Similarly, the LSA model performs dimensionality reduction by applying singular value decomposition (SVD) on the representation matrix  $N$  as in Figure 3.1b. Matrix  $\Sigma$  contains the sorted  $r$  singular values in its main diagonal. Thus, LSA considers the  $k$  latent topics the  $k \leq r$  most significant singular values by setting the remaining in the matrix to zero.

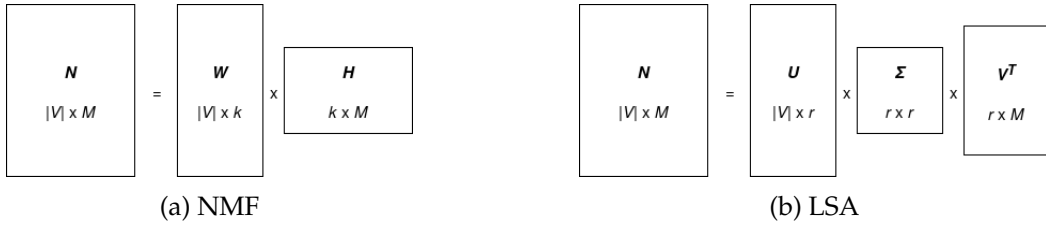


Figure 3.1: Decomposition of the term-document collection representation matrix  $N$  following different approaches.

On the other hand, probabilistic approaches assume a generative probabilistic process for each document. The probabilistic variation of LSA (pLSA) (Hofmann, 1999) performs decomposition over mixtures of multinomial components, sampled from an estimated latent variable “aspect” model. The aspect model estimates the mixture  $P(d, w) = P(d)P(w|d)$ , for each term and document, where  $w$  is independent from  $d$ , given an observed class variable  $z_i \in Z$ , as defined in Eq. (3.2).

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \quad (3.2)$$

The Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) model represents each document as a mixture of multinomial distributions (defined in the latent topic space), in which each multinomial is defined over the vocabulary space. Each topic mixture  $\theta$  is sampled from a  $k$ -dimensional Dirichlet distribution (the number of topics  $k$  is defined a priori), parameterized by  $\alpha$  which intuitively models the concentration of topics per document in a collection. Finally, each multinomial is parameterized by another Dirichlet prior  $\beta$ , which models the concentration of words per topic. This relaxed paradigm allows for many-to-many relationships both between topics and words, and documents and topics, which fits the intuition that a document may comprise several topics and that a word may belong to multiple topics.



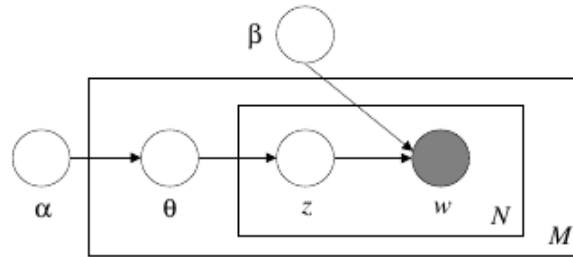


Figure 3.2: LDA box diagram, where each box is a replicate. The outer box represents the collection of documents and the inner box the sequence of topics and words per document. Figure adapted from Blei et al. (2003).

### 3.1.2 Topic Modeling: Evaluation metrics

The performance of topic models may be intrinsically evaluated regarding topic coherence through mutual information and perplexity, given that the model provides a distribution over the vocabulary, which is the case for the probabilistic approaches. Topic coherence measures how semantically related are the top words of a given topic, and averaging over all topics yields the model's coherence. Specifically, the Pairwise point-wise Mutual Information (PMI) score, defined by Eq. (3.3), gives a higher score to topics which  $T$  top words are more likely to co-occur in the same document, normalized against their individual independent probability in the collection. This measure is said to account for topic coherence because it encodes the notion that words defining a concept, that often share the same context, "gain" in information from one another to provide with a more well-defined, or coherent, topic. This metric is dependent on the used corpus and therefore carries any statistical lack of information that might exist in said corpus, for instance, considering a collection of documents with a lack of word co-occurrence information, this will negatively impact the PMI score, if it is indeed calculated on that collection. In these cases, a possible way to circumvent this problem is to evaluate the resulting topics with the PMI score on external collections which do not have that lack of information. Topic coherence may also be measured by expert evaluation, but this approach is usually not considered due to the expense of using human judges.

$$\text{PMI}(t) = \frac{2}{T(T-1)} \sum_{i < j \leq T} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (3.3)$$

Topic models that follow probabilistic approaches estimate mixtures over the latent se-

semantic space and a distribution over the vocabulary. These distributions may be evaluated regarding how well they model never-seen data. A model’s perplexity intuitively measures the inverse likelihood of the test data, so that the better it fits the model, the lower perplexity score is obtained (the less “perplexed” the model is to the new data). Formally, the perplexity of a model given a test set is given by Eq. (3.4) (as defined by [Blei et al. \(2003\)](#)). However, perplexity has been shown to not reflect semantic coherence of a topic, sometimes scoring against expert evaluation ([Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009](#)).

$$P(D_{test}) = \exp \left( - \frac{\sum_{m \in D_{test}} \log p(m)}{\sum_{m \in D_{test}} |m|} \right) \quad (3.4)$$

### 3.1.3 Short-text topic modeling

In certain contexts, there is a useful focus on short-text, particularly due to the necessity of analyzing data derived from online platforms such as social media (e.g. Twitter posts). Extracting topics from short texts, where the document length has shifted from the hundreds of words to the hundreds of characters, presents challenges that the traditional models are not capable of efficiently overcoming, specifically the difficulty in capturing word co-occurrence information, due to noise and sparsity. This has led to a recent line of research which has introduced enhanced traditional models with external semantic representations and term correlation. The motivation is two-fold: (i) external semantic representations provide a good partitioning of the semantic space, clustering together words that are related in a given context; (ii) external semantic representations can be derived from larger datasets which may not have the restrictions identified in short-text documents.

However, as pointed out by [Sridhar \(2019\)](#), the partitioning of the semantic space is dependent on the task at hand, for instance, if a classification approach were to be followed, the terms clustered together are expected to be that of each class in the problem, but on the other hand, following the language modeling approach, terms that often share the same context (surrounding window of terms) are expected to be clustered together, which for generic tasks of topic modeling is the desired behavior.

The Biterm Topic Model (BTM) ([Yan, Guo, Lan, & Cheng, 2013](#)) is a probabilistic approach that tackles sparsity in short-text document collections by virtually aggregating the whole collection into a single, long document. Thus, instead of capturing word co-occurrence at the

document level, this information is captured at the collection level, so that it describes a generative process of word co-occurrence patterns instead of documents. Co-occurring words, denominated biterms, are modeled to be sampled from a topic multinomial  $\rho_z$ , which in turn is sampled from a Dirichlet topic distribution  $\theta$  of the whole collection. The parameters  $\rho$  and  $\theta$  are estimated similarly to LDA. Even though BTM has been shown to outperform LDA regarding topic coherence, it does not make use of external semantic information and thus is limited to the statistical information present in the training short-text collection.

The Latent Feature LDA (LF-LDA) model (Nguyen, Billingsley, Du, & Johnson, 2015), in contrast, incorporates latent features (semantic representations) learned from external corpora in the LDA model, specifically in the topic-to-word Dirichlet multinomial component. Thus, it assumes a generative process similar to that of LDA with the exception that, for each word, a binary indicator is sampled from a Bernoulli distribution in order to decide if the word should be generated by the multinomial topic distribution or the latent feature (additional) component. This allows for the external semantic information to model topic-to-word generation when there is lack of information in the documents themselves. The external models used to export the latent features, such as Word2Vec and skip-gram (Mikolov, Chen, Corrado, & Dean, 2013), are neural-network-based models which learn real-valued vectors that represent words in a language modeling semantic space. These vectors, denominated word embeddings, are dense and of arbitrary length  $L$ , usually chosen in such a way that  $L \ll |V|$ , so that document representation may be done in a lower-dimensionality space. FastText (Mikolov, Grave, Bojanowski, Puhersch, & Joulin, 2017) is another such model, with the difference that it learns embeddings at the character level, virtually infinitely expanding the accepted vocabulary space (considering that any word is a combination of characters and that its corresponding embedding is a combination of the character embeddings).

The Generalized Pólya Urn Dirichlet Multinomial Mixture (GPU-DMM) (Li, Wang, Zhang, Sun, & Ma, 2016) follows a similar approach, but instead of directly incorporating the external semantic information into the generative process, during the estimation step, promotes semantically related words to be assigned to the same topic. The generative process on top of which it is built is given by the Dirichlet Multinomial Mixture (DMM) (Yin & Wang, 2014) model that makes the simplifying assumption that each short text document talks about one single topic, so that all words in a document are sampled independently from the same topic multinomial

distribution, which in turn is sampled from the topic Dirichlet mixture of the collection. Both LF-LDA and GPU-DMM have been shown to have higher performance regarding topic coherence when compared against LDA and DMM baselines, effectively demonstrating the gain of incorporating external semantic information into the topic model.

In short texts, LDA and LDA-based models have been shown to under-perform when compared against NMF and NMF-based models with respect to the PMI topic coherence evaluation metric (Y. Chen, Zhang, Liu, Ye, & Lin, 2019), which is argued to be due to the sparsity and noise of short texts, the instability of stochastic Gibbs sampling when there is not sufficient term co-occurrence information, and the fact that NMF can operate in matrix representations of collections which might encode term discriminative information, such as the TFIDF representation matrix. For these reasons, current research has focused on short-text topic modeling following non-probabilistic approaches, specifically with NMF. These are presented below.

The knowledge-guided NMF (KGNMF) (Y. Chen et al., 2019) adds a semantic constraint to the factor matrices  $W$  and  $H$ , which states that word representation in the latent topic space (rows of  $W$ ) should preserve the relatedness between word-pairs in the external semantic representation space. This word-word semantic graph regularization matrix  $S$  is learned from an external corpus, where each entry  $s_{ij}$  is the cosine similarity between two word vectors, defined by Eq. (3.5). These word vectors are specifically learned with Word2Vec. Thus, whilst maintaining the non-negative constraints, NMF is adapted so that Eq. (3.6) is minimized. Even though precise values are not provided by the authors (only comparative graphs, lacking specificity), KGNMF is shown to have better performance with respect to the PMI score, when compared against BTM, LF-LDA, LF-DMM (Nguyen et al., 2015), GPU-DMM, GK-LDA (Z. Chen et al., 2013), LDA, and NMF, especially when considering the top 5 and 10 words of each topic. Comparing with the NMF baseline exposes the gain of external semantic information enhancement.

$$s_{ij} = \cos(w_i, w_j) = \frac{\langle w_i, w_j \rangle}{\|w_i\| \times \|w_j\|} \quad (3.5)$$

$$\|N - WH\|_F^2 + \lambda \times \text{tr}(W^T LW), \quad (3.6)$$

where  $L = \text{diag}(S \times \mathbb{1}) - S$

$$\text{tr}(W^T LW) = \min \frac{1}{2} \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} s_{ij} \|w_{i*} - w_{j*}\|_2^2 \quad (3.7)$$

Viegas et al. (2018) propose a semantically enhanced NMF model that replaces each word in a document by the corresponding embedding obtained from the Word2Vec model. Each document is then represented by a single vector obtained from the Fisher Vector of single multivariate Gaussian distribution pooling strategy, applied over the corresponding document word embeddings. This is a simplified version of the Fisher Vector pooling strategy (FV) (Lev, Klein, & Wolf, 2015). The decomposition step is applied as defined in Figure 1, however the explicit relation between topics and words is effectively lost as matrix  $W$  columns represent latent topic weights over the FV space instead of the vocabulary space (noting that matrix  $H$  still represents linear combinations of topics for each document). In order to overcome this, a novel topic extraction step denominated Advanced Semantic Topic Combination (ASToC) is proposed. This strategy introduces a way to group documents represented in FV dimensions into the decomposed topics, and the use of Information Gain (IG) to associate the most relevant words to each topic. In summary, this strategy builds a tripartite graph with three types of nodes, each corresponding to the documents, latent factors (topics), and FV features or dimensions, where each edge represents the relation between the nodes, weighted by the probability distribution derived from the factor matrices, such that document nodes and topic nodes are weighted by matrix  $H$  and topic nodes and FV nodes are weighted by matrix  $W$ . This graph is then submitted to a topic merging step to increase topic coherence whilst updating matrix  $H$  to reflect the merge changes, so that if topic  $k_i$  is merged with  $k_j$ , then rows  $H_i$  and  $H_j$  are removed from the matrix and row  $H_*$ , as defined by Eq. (3.8), is appended to the bottom. The final step applies the IG technique to the set of documents of each latent topic in order to obtain the most relevant words associated with that topic. This model is demonstrated to have significant improvements over BTM, LDA, GPU-DMM, Feature Sentiment (FS) (Guzman & Maalej, 2014), Life-long Topic Model (LTM) (Z. Chen & Liu, 2014), Embedding-based Topic Model (ETM) (Qiang, Chen, Wang, & Wu, 2017), and Additive Regularization of Topic Models (ARTM) (Vorontsov & Potapenko, 2015), with respect to the normalized PMI (NPMI) score (defined in Eq. (3.9)), in all twelve analyzed datasets. From the extensive experimental results, it is possible to observe that the proposed model has a higher NPMI score in the vast majority of cases (36 out of 49), in some having over 50% increase over the second best baseline, for

instance, considering the top  $T = 10$  words of each topic, the second highest scoring model is the GPU-DMM with a score of  $0.304 \pm 0.157$ , whilst the proposed model scores  $0.620 \pm 0.055$ .

$$H_* = \frac{H_i + H_j}{2} \quad (3.8)$$

$$\text{NPMI}(t) = \sum_{i < j \leq T} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-p(w_i, w_j)} \quad (3.9)$$

The semantics-assisted NMF (SeaNMF) (Shi, Kang, Choo, & Reddy, 2018) model overcomes the problems associated with short-text noise and sparsity by applying a skip-gram model with negative sampling (SGNS) with a context window size equal to that of each document (given that it is applied to short texts). The skip-gram model is used because it learns to predict a context window (set of surrounding words) given a single word from the vocabulary, effectively learning a word vector  $\vec{w}_i \in \mathbb{R}_+^k$  and a context vector  $\vec{c}_j \in \mathbb{R}_+^k$  for each  $w_i, c_j \in V$ . By constraining these vectors to be non-negative, matrix  $W$  (Figure 1) is defined so that  $W(i, :) = \vec{w}_i$  and corresponding context matrix  $W_c(j, :) = \vec{c}_j$ . Thus, the term-context correlation matrix  $S$  is obtained by  $S \approx WW_c^T$ . This strategy is shown to capture relevant term-context correlation that otherwise would not be fully taken advantage of by the traditional NMF model. At this point, a bi-relational collection representation matrix with both term-document and term-context information is obtained by vertically stacking  $N^T$  and  $S^T$ . Finally, the objective function, defined by Eq. (3.10), where  $\alpha \in \mathbb{R}_+$  is a scale parameter and  $\psi(W, W_c, H)$  is a penalty function specified for sparsity, is solved using a block coordinate descent algorithm. Given the previously mentioned problem of the PMI evaluation for short texts, the experimental results of SeaNMF include the PMI scores, both against the short-text datasets used for training, but also on external datasets composed of long documents, specifically the Yahoo.CA (Research, accessed December 30, 2019) and ACM.IS (Luo, 2014 (accessed December 30, 2019)) datasets. SeaNMF is compared against LDA, NMF, GPU-DMM, and Pseudo-document-based Topic Model (PTM) (Zuo et al., 2016), consistently showing higher scores, in some cases with a performance score (3.6318) over 50% higher than the second best model (PTM with 1.6628). SeaNMF is shown to outperform NMF, which evidences the need for taking advantage of semantic information when considering short texts. On top of this, it is argued that the fact that the semantic information is learned from the collection itself (and not from an

external source, such as the case of GPU-DMM) is a determining factor due to the possibility of introducing bias from context-inadequate semantic spaces.

$$\min_{W, W_c, H \geq 0} \left\| \begin{bmatrix} N^T \\ \sqrt{\alpha} S^T \end{bmatrix} - \begin{bmatrix} H \\ \sqrt{\alpha} W_c \end{bmatrix} W^T \right\|_F^2 + \psi(W, W_c, H) \quad (3.10)$$

The cluster-of-words (CluWords) (Viegas et al., 2019) model exploits external semantic information by replacing each term in a document bag-of-words (BOW) representation by a meta-word, denominated CluWord, which represents the cluster of syntactically and semantically similar words. Each term’s CluWord  $C_t$  is a row in the CluWords matrix  $C_{|V| \times |V|}$ , where each entry  $c_{t,t'}$  is the cosine similarity score between the pre-trained word embedding of term  $t$  and term  $t'$ ,  $\forall t, t' \in V$  (scores below a threshold  $\alpha$  are set to zero). For this extended BOW representation to be fully taken advantage of, the model incorporates a TFIDF-based approach capable of weighting the semantic information carried in each CluWord, defined by Eq. (3.11). In this approach, matrix  $C_{tf} M \times |V|$  represents the term frequencies of each CluWord in each document, so that row  $C_{tf,d}$  is given by the sum of the products of the frequency of each term  $t$  in document  $d$ , given by  $T_{d,t}$ , and the corresponding similarity measure in the CluWord given by  $C_{t,t'}$ , as defined in Eq. (3.12). Matrix  $\text{idf}(C)$  determines the inverse document frequency of each CluWord  $C_t \in C$  as defined in Eq. (3.13). The term  $\mu_{C_t,d}$  is the mean of the values of the similarities in CluWord  $C_t$  that occur in the vocabulary sub-set of all terms in document  $d$  which have similarity not equal to zero in  $C_t$ . The novel TFIDF-based CluWord representation matrix  $C_{\text{tfidf}}$  is then submitted to factorization as in the traditional NMF model.

The experimental results show that CluWords derived from the pre-trained word embedding FastText model trained with WikiNews (Mikolov et al., 2017) achieves better NPMI scores when compared with Word2Vec trained with GoogleNews (Mikolov et al., 2013) and FastText trained with Common Crawl (Mikolov et al., 2017), but the results were mainly statistical ties, which suggests that the CluWords model is capable of extracting equally coherent topics in all three semantic spaces (i.e., to some degree, it is capable of avoiding the bias in the pre-trained models). With threshold  $\alpha$  chosen so that 2% of the most similar words are selected with the pre-trained word embedding FastText WikiNews model, when compared against FS, BTM, LDA, LTM, GPU-DMM, ETM, ARTM, and SeaNMF, in twelve different short-text datasets, CluWords achieved the best NPMI score seven times and statistically tied the remaining exper-

iments with SeaNMF. However, taking into account the standard deviation of the scores, the ones obtained by CluWords are considerably smaller than those of SeaNMF, and are indeed less variable according to the [Levene \(1960\)](#) and [Bartlett \(1937\)](#) variability tests.

$$C_{\text{tfidf}} = C_{\text{tf}} \times \text{idf}(C) \quad (3.11)$$

$$C_{\text{tf}} = T \times C \quad (3.12)$$

$$\text{idf}(C) = \log \frac{M}{\sum_{1 \leq d \leq M} \mu_{C_t,d}} \quad (3.13)$$

## 3.2 Audio-based analysis

In addition to conveying the words used in linguistic analysis, the speech audio signal may also carry relevant information about the speaker. This can then be used to further characterize what is being said or infer states about the speaker. This type of information is useful in many cases. The tasks of speech recognition and identification of speaker's intention use this information to further enrich the linguistic analysis, for instance, with disambiguation. The diagnosis of speech disorders is directly linked with the extraction of high level speech features, such as reduction of spontaneous verbalizations, trouble finding words, and degradation of articulation. Systems that include human-computer interfaces are also usually interested in the detection of anger and stress levels, for instance, in speech interactions in call centers, and in the detection of other states, such as uncertainty, interest, and deception, all of which can be used to adapt these interfaces or any other embedding system ([Schuller et al., 2013](#)).

In the present work, it is of interest to further complement the textual content of the descriptions of pain with information about the patient, namely, the emotional state, affection, stress, insecurity, tiredness, irony, sarcasm, and other speech characteristics, which can, in some way, affect the perception of pain or intrinsically reveal characteristics of the painful experience. The following sections present the signal features which are believed to carry relevant information, the general architecture of information extraction models from speech, how they



can be evaluated, and, finally, focus on the task of speech information extraction specifically for emotion recognition.

### 3.2.1 Frame and utterance level features

It is largely assumed that the relevant information in audio-based speech analysis is indeed in its temporal variation rather than static values. To this end, acoustic features are captured at the frame level (20-50 ms), which are then statistically aggregated over the duration of the signal, or utterance, in the specific case of speech, in order to obtain variational utterance level features.

Commonly extracted frame level features, denominated Low Level Descriptors (LLDs), include the pitch (through the  $F_0$  fundamental frequency), which allows for the auditive distinction between lower and higher sounds, the energy, defined as the area below the squared magnitude of the continuous-time signal (which in this case is discretized at the specified sampling rate), which allows for the auditive distinction between quieter and louder sounds, the zero-crossing rate, defined as the rate at which the sign of the speech signal changes, which intuitively carries information about the smoothness of the sound, and the Mel Frequency Cepstral Coefficients (MFCCs), which are a set of features that concisely describe the sequence of frames, by taking the constituent frequencies of the signal with a Fourier transform, mapping them onto the Mel-frequency cepstrum (MFC), since evenly spaced frequencies in the MFC are closely related to the human hearing, and finally taking the corresponding power logs, which is also motivated by the human hearing, which is less capable of distinguishing variations in higher frequencies than lower frequencies. The MFCCs are the amplitudes resulting from taking the discrete cosine transform of these Mel-frequency log powers.

In order to reveal variation over time (during the duration of the signal), these LLDs are aggregated using statistical functions, which comprise the identification of maximum and minimum values, mean, standard deviation, range, and high-order derivatives, effectively characterizing the signal at the utterance level. Higher level paralinguistic features may also be extracted, such as laughter, sighs, and disfluencies, including hesitation, repetitions, and pauses (Schuller et al., 2013).

### 3.2.2 General architecture and evaluation

Speech models, in general, follow a common architecture, in part because all tasks deal with similar challenges intrinsic to the speech signal. On the first stage, the signal is preprocessed, so that noise is removed. If there are multiple speakers, these are separated through speaker diarization, and the signal is further separated, for instance, using NMF, if deemed necessary. Next, acoustic features are extracted at the frame level (the LLDs). These describe the signal statically, at each sample. Most tasks are interested in the temporal variation of the signal, rather than static information. To this end, a set of statistical aggregation functions are applied to each LLD, over the duration of the signal. At this stage, utterance level features are obtained. Large feature vectors are then usually submitted to dimensionality reduction, for instance, using PCA, before being used for training. The machine learning model used will depend on the task at hand. For classification purposes, Support Vector Machines (SVM) and Hidden Markov Models (HMM) are typically employed.

When dealing with a classification problem, which is usually the case for Speech Emotion Recognition (SER), the evaluation metrics that are commonly used are the unweighted accuracy, weighted accuracy, and  $F_1$  score. These are usually employed in the multi-class setting, if there are highly unbalanced test classes. Calculating the regular accuracy could yield false conclusions, since the most representative classes would statistically dominate the metric.

The unweighted accuracy (UA) is given by the average recall of each class, as defined by Eq. (3.14), where  $TP_c$  refers to the true positives,  $TN_c$  the true negatives,  $FP_c$  the false positives, and  $FN_c$  the false negatives, all referring to class  $c$ .

$$UA = \frac{1}{|C|} \sum_{c \in C} \frac{TP_c}{TP_c + FN_c} \quad (3.14)$$

The weighted accuracy (WA) metric, specifically the uniform weighted accuracy, defined by Eq. (3.15), gives equal contribution to the resulting metric to each class, independent of the number of samples.

$$WA = \frac{1}{|C|} \sum_{c \in C} acc(c) \quad (3.15)$$

$$\text{acc}(c) = \frac{\text{TP}_c + \text{TN}_c}{\text{TP}_c + \text{FP}_c + \text{TN}_c + \text{FN}_c} \quad (3.16)$$

The  $F_\beta$  score measures the accuracy of a given model by having into account both the precision (P) and recall (R) metrics. The  $F_\beta$  score for a single class is given by Eq. (3.17), where  $\beta$  is the degree of importance of the recall metric in relation to the precision metric. The  $F_1$  ( $\beta = 1$ ) score is the harmonic mean of both precision and recall, so that its value tends to the smaller of the two. Thus, the best  $F_1$  score (equal to one) is obtained when both precision and recall are perfect as well.

$$F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times R) + P} \quad (3.17)$$

### 3.2.3 Speech emotion recognition

Speech information extraction consists on the extraction of the aforementioned features present in the audio signal, in order to study their relations, both between themselves and other external information. Emotion has been identified as a key aspect in the perception and expression of pain. Thus, the present work will initially focus on the task of SER, so that this type of models may be used to characterize the signal.

SER is the paralinguistic task which focuses on recognizing emotion given a speech audio signal. The output can vary from categorical variables (sad, happy, angry, surprised, etc.) to continuous values in specific dimensions, for instance, valence and arousal. Data points of either type can be mapped to the other.

Handcrafted feature-based approaches extract the utterance level features by statistically aggregating the LLDs, and then apply machine learning algorithms, such as SVM (for the classification case). Recent developments, however, have focused on automatically learning both frame and utterance level features, typically following the artificial neural network approach. The reasoning is two-fold (Mirsamadi, Barsoum, & Zhang, 2017): (i) it is unclear which features best characterize the signal for emotion recognition; (ii) variational analysis depends on pre-processing the signal in order to remove frames which are known not to carry emotion information, specifically silence frames, and it should also employ a weight-based analysis since often only a few parts of the signal carry relevant information, rather than the whole utterance.

Ignoring the latter could imply the distortion of the decision making due to noise in the feature set. This has led to the development of end-to-end neural-network-based architectures, that take as input the raw temporal data samples and output an emotion label or value, which have shown improvements over the traditional approach. Commonly employed structures are specializations of Deep neural networks (DNN), including Recurrent neural networks (RNN), specifically, Long short-term memory networks (LSTM), and Convolutional neural networks (CNN), which are briefly presented below to contextualize the state-of-art.

LSTMs are RNNs specifically modeled to process sequences of inputs, such that the computation of the input at time-step  $t$  is dependent of previous time-steps, by maintaining an internal state (memory). In this architecture, the memory of past computations is modeled by other networks, which in turn integrate gating mechanisms that determine which of the previous information is relevant for the current computation. A LSTM unit is composed of a cell, which holds actual memory of previous computed values, and, commonly, three gates, the input, forget, and output gates. In summary, these gates are responsible for determining if a value should be stored in the cell, if a value should remain in the cell, and if the value in the cell should be used to compute the activation of the unit, respectively. This design renders LSTMs well-suited for the processing of temporal sequences of inputs, due to the relative insensitivity to the dependence distribution of information in the whole sequence.

CNNs are a specialization of DNNs in which at least one of the hidden layers performs the mathematical operation of convolution. In practice, this operation is a sliding dot product of the input window and a given filter, producing an activation map of that filter, which is then passed onto the next layer. During training, these filters are learned, so that they are only activated in the presence of specific features, or patterns, in the input, thus, stacking more convolutional layers allows for the detection of increasingly more complex patterns in the data. In addition, these layers are often interleaved with pooling layers, which perform down-sampling of their input, in order to reduce the number of parameters and avoid overfitting. Max pooling is a common pooling strategy, where only the maximum value of the input window is passed onto the next layer. This type of dimensionality reduction is motivated by the fact that the relative spatial position between features is often more important than the exact location of each one.

[Mirsamadi et al. \(2017\)](#) propose an RNN-based model, which architecture consists of two main parts (Figure 3.3). The first part is composed of two feed-forward layers, each with 512

neurons and Rectilinear Unit (ReLU) activation, defined as  $h(x) = \max(0, x)$  where  $x$  is the neuron's net value, and is responsible for learning the frame level LLDs. The second part, composed of two bidirectional LSTM networks, which connects layers in both directions, so that information from past and future computations is taken advantage of, with a weighted pooling attention-based mechanism, is responsible for learning the utterance level, temporal, and statistical aggregation of each LLD. The weighted pooling layer, placed at the end of the architecture, is based on an attention mechanism, in this case composed of a Logistic Regression, capable of learning which frames of the input carry the most relevant information for the predicted output, by associating an adequate weight to each frame. In this way, the pooling layer can use these weights to take advantage of relevant frames and ignore the noise in the remaining. The experimental study, on the IEMOCAP dataset (Busso et al., 2008), considering only audio signals from four emotion classes, revealed that the end-to-end architecture, even though it had WA and UA scores slightly higher (+4.0% and +0.6%) than the baseline (57.8% and 55.7%), which is a Support Vector Machine (SVM) classifier based on handcrafted LLDs and utterance level statistical features, did not have a much better performance due to the lack of data for the task's complexity. However, the hybrid architecture, with handcrafted LLD features and the LSTMs with weighted pooling mechanism, revealed the advantage of using attention mechanisms to learn frame importance, with gains of +5.7% and +3.1% in WA and UA respectively, over the baseline.

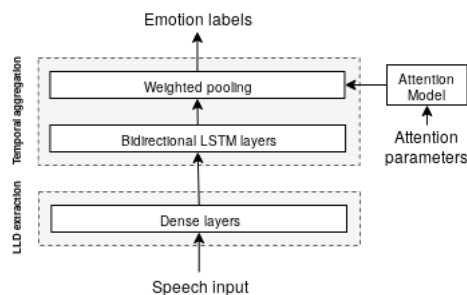


Figure 3.3: Attention-based weighted pooling NN. Adapted from Mirsamadi et al. (2017).

Another possible approach for emotion recognition is to take into account information captured from multiple domains, for instance, text and speech, since emotions are indeed expressed through multiple channels. Thus, fusion techniques take advantage of intra- and inter-domain dynamics for emotion recognition. Early fusion concatenates features extracted from multiple domains and feeds them to a single model. This approach is able to capture inter-

domain dynamics, but may fail to capture the variability within a single domain, as it attempts to model, in general, all the domains' features. Late fusion, on the other hand, only merges information before making the classification/regression decision. To this end, this approach has one model for each domain, effectively capturing domain specific variability, and uses a voting decision system between all models to make the final prediction. However, this comes at the risk of not capturing inter-domain dynamics when the feature spaces are very different.

Recent research has, thus, led to the development of emotion recognition models based on the aforementioned fusion approaches. Specifically, [Sebastian and Pierucci \(2019\)](#) propose a deep learning architecture that combines both intra- and inter-domain dynamics, particularly the text and speech domains, with both early and late fusion. In summary, the architecture can be split into three main components (see [Figure 3.4](#)). The feature extraction component extracts both text and speech features from the input: text features are extracted from pre-trained 300-dimensional word embeddings from FastText, using convolutional, max pooling, and fully connected layers; speech features are extracted using the OpenSMILE toolkit ([Eyben, Wöllmer, & Schuller, 2010](#)), and comprise both LLDs and utterance level speech features. The text features are fed into the uni-domain (textual) component, which performs emotion recognition solely based on text. This component comprises a LSTM layer, two fully connected layers, and, finally, an output layer with as many units as emotion classes. The text features are also concatenated with the speech features using early fusion, and fed to a joint component, which accounts for inter-domain dynamics. This last component comprises two CNN layers with max pooling, followed by two fully connected layers, and the output layer. This component was chosen not to be based on LSTMs because it showed poorer performance compared to the CNN-based approach. Finally, the outputs of the uni-domain and joint models are merged using late fusion to output the final prediction.

The experimental setup includes the study, on the IEMOCAP dataset with six emotion classes, of multiple late fusion techniques, showing weighted voting and product rule combination, defined by [Eq. \(3.18\)](#) and [Eq. \(3.19\)](#), respectively, to have the best performance regarding UA, WA, and  $F_1$  scores. Specifically, the latter having 60.2%, 61.2%, and 61.2%, respectively. The baselines used for comparison against the best late fusion combinations, include the Tensor Fusion Network (TFN) ([Zadeh, Chen, Poria, Cambria, & Morency, 2017](#)), Memory Fusion Network (MFN) ([Zadeh et al., 2018](#)), Bi-directional contextual LSTM (cLSTM) ([Poria et al., 2017](#)),

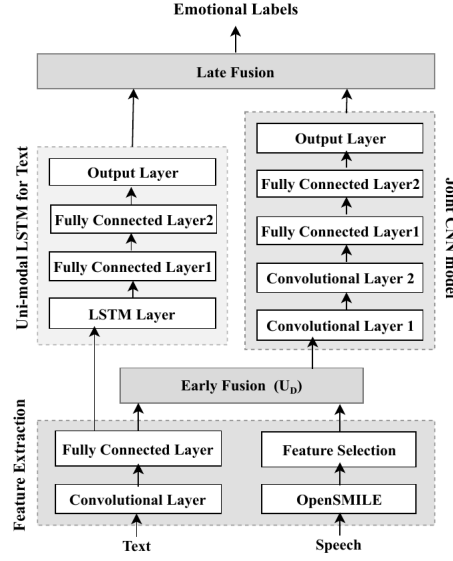


Figure 3.4: Block diagram of an emotion recognition model architecture, based on speech and transcripts, integrating both early and late fusion. Adapted from [Sebastian and Pierucci \(2019\)](#).

and Interactive Conversational Memory Network (ICON) ([Hazarika, Poria, Mihalcea, Cambria, & Zimmermann, 2018](#)), noting that the last two models are dialog level emotion recognizers, which carry more contextual information for the decision making than utterance level recognizers. Again, the proposed architecture, combined either with weighted voting or product rule late fusion, showed the best performance results regarding UA, WA, and  $F_1$  scores, excluding the dialog based emotion recognizers. The second best performing baseline was MFN (which also uses video-based features), scoring, respectively, 58.3%, 60.1%, and 59.9%, whilst the proposed architecture had gains of +1.9%, +1.1%, and +1.3%. The performance of the ICON model was also compared to show that utterance level emotion recognition based on the proposed architecture can compete with dialog level recognition, having only close to a -2% decrease in performance.

$$s = w_1 \times s_{\text{text}} + w_2 \times s_{\text{text,speech}} \quad (3.18)$$

$$s_j = \frac{s_{j_{\text{text}}} \times s_{j_{\text{text,speech}}}}{\sum_i s_{i_{\text{text}}} \times s_{i_{\text{text,speech}}}}, i \neq j, i, j \in C \quad (3.19)$$

### 3.3 *Summary*

In this chapter we explored the literature related to the analysis of language, specifically for the language of pain.

Starting with the text-based analysis, we focused on the technique of topic modeling. We exposed the traditional methods, specifically the probabilistic and non-probabilistic approaches (e.g. LDA and NMF, respectively). We identified the challenges associated with our data, namely its short-text nature, and concluded on the necessity of tailored models to tackle its intrinsic noise and sparsity. We presented an extensive review of short-text topic models and identified the deficiencies of probability-based approaches in this setting. Thus, we mainly focused on the review of the methods based on the non-probabilistic approach, specifically NMF, with the introduction of semantics assistance, such as word embeddings.

Finally, for the audio-based analysis we presented the traditional approach to information extraction from speech, including the generic architecture, the commonly extracted feature set, and the set of evaluation metrics commonly employed. With a focus on the task of Speech Emotion Recognition, as an instance of Speech Information Extraction, we explored how this task is tackled in the literature. Specifically, we explored more recent end-to-end models based on neural-networks that merge features from multiple domains (e.g. text and speech).



# 4 Dataset Definition

In this chapter we define the dataset used in the present study. All data were collected and prepared with the objectives previously presented in mind. This dataset is the result of a joint data collection project with the Faculty of Medicine of University of Porto (FMUP), which took place at University Hospital Center of São João (UHCSJ), for a total of twelve months (from October, 2019, to October, 2020). The data includes verbal descriptions of chronic pain experiences (resulting from recorded, scripted interviews) and additional contextual information (demographic and clinical data) from patients deemed eligible for the study. These are adults (older than or equal to eighteen years of age), of either sex, diagnosed with osteoarthritis, rheumatoid arthritis, or spondyloarthritis (including psoriatic arthritis), and with symptoms of chronic pain. A total of 94 patients were included in the collection.

All data were collected under a collection protocol, approved by the Ethics Commission of UHCSJ, in which data confidentiality is explicitly protected. All recordings are anonymous, and the presentation of results is always made without individual references. Patient recordings are identified with a unique ID, and kept separate from the ID resolution key, which is maintained in physical format at a secure location. The ID also links recordings with other relevant data, so that the patient personal identification is never used.

## 4.1 *Collected information*

For each patient, we collect a natural description of the personal experience of chronic pain and contextual demographic and clinical parameters of disease and pain manifestations.

The description of the experience, on the one hand, should be based on the aspects of the experience that are most important to the subject describing it, and must be based on that subject's elicited vocabulary alone. On the other hand, as suggested by the presented literature, there are specific cognitive aspects that have been identified as the most determinant factors

in assessing experiences of pain, therefore, each subject should provide their insights for each of these cognitive aspects, with less or more importance, depending on their personal experience. These natural descriptions should, therefore, be unrestricted in terms of the described aspects, but at the same time guided to the previously identified cognitive aspects. This may be achieved with a specifically designed interview, with open, guiding questions. Finally, as stated before, the context in which pain is described influences both the vocabulary used and the way it is verbalized, thus, the description is obtained in the medical office, immediately after the regularly scheduled appointment, by the health professional.

The set of questions composing the interview was the result of a design process that aimed at obtaining a natural description of the patient's pain experience, in their own words, but, at the same time, directing it towards the cognitive topics that were identified as the most relevant for pain assessment. The script, validated by multiple health professionals included in the collection process, is as follows (translated from Portuguese):

1. *Onde localiza a sua dor?*

Where does it hurt?

2. *Como descreveria a sua dor? Como a sente/que sensações provoca?*

How would you describe your pain? How do you feel it/which sensations does it cause?

3. *Como tem evoluído a intensidade da dor no último mês?*

How has pain intensity evolved in the past month?

4. *Como considera que a dor tem afetado o seu dia-a-dia, nomeadamente na sua atividade física, profissional e social, e o seu estado emocional?*

How would you consider pain to affect your day-to-day, namely, your physical, professional, and social activities and your emotional state?

5. *Qual considera ser a origem da sua dor?*

What do you believe to be the cause of your pain?

6. *Como considera que tem evoluído a sua dor, tendo em conta o tratamento (atual) aplicado?*

How would you say your pain has evolved, considering the current treatment?

7. *Como acha que irá evoluir a sua dor nos próximos meses?*

How do you expect your pain to develop in the coming months?

The contextual information is comprised of basic demographic information (age, gender, and education level), duration of the disease and reports of pain, the therapeutic processes, analytical parameters of the disease's activity (Erythrocyte Sedimentation Rate (ESR) and Repetitive C-Protein (RCP)), and VAS of both pain and disease. The form used to collect these data is shown in Figure 4.1.

Posterior to this collection, each patient recording was listened to and evaluated by a health professional (which did not participate in the collection and has no known connections with the patient). Indeed, some patients were evaluated by multiple health professionals, but not all. This evaluation produces an additional perspective on the patients in terms of linguistics, since it is performed by professionals and is limited to the audio alone. The evaluation consists of comments on the situation of the patient regarding pain and disease management, emotional state, and patient perception of the situation. Additionally, segments of the patient speech were identified as evidence for the provided evaluation.

## Formulário

ID: \_\_\_\_\_

## Informação demográfica básica

Idade: \_\_\_\_\_

Género:  Masculino  Feminino  Outro: \_\_\_\_\_

Grau de escolaridade: \_\_\_\_\_

Ocupação profissional: \_\_\_\_\_  Activa  Baixa/Reforma

## Patologia

 Osteoartrose  Artrite reumatoide  
 Espondiloartrites  Outra: \_\_\_\_\_
EVA (dor): \_\_\_\_\_  
 Sem dor  Dor máximaEVA (doença): \_\_\_\_\_  
 Muito bem  Pior possível

Tempo de evolução das queixas algícas: \_\_\_\_\_

Duração da doença: \_\_\_\_\_

Terapêuticas instituídas:


Parâmetros analíticos da atividade da doença

VS: \_\_\_\_\_

PCR: \_\_\_\_\_

Figure 4.1: Form used to collect the contextual information of each patient. The form is filled by the health professional with the assistance of the patient.

## 4.2 Data preparation

In order for the collected data to be processed in a systematic and automatic way, the raw data of each patient is put through a preparation pipeline, depicted in Figure 4.2.

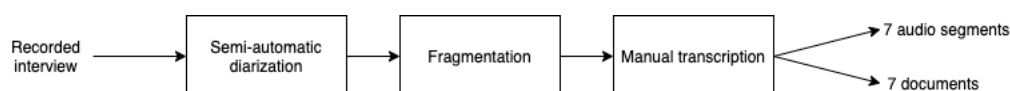


Figure 4.2: Preparation pipeline of the data, which is used to create the baseline dataset.

Given as input an audio file with a recorded interview, the first stage of the pipeline is speaker diarization, which comprises the segmentation of the audio file by speaker, so that in each segment there is only one identified speaker. It is assumed that during the interview only two subjects speak, the interviewer and the interviewee. The implementation is detailed after. The second stage is the fragmentation of the audio file by question in the interview, resulting in a total of 7 segments per patient. These segments include only the interviewee’s speech. Finally, each of these fragments is manually transcribed. The strategy comprises a clean transcription, which does not account for repetitions, corrections, hesitations and other speech disfluencies. At the end of the pipeline, to each patient is associated a set of 7 audio segments and corresponding 7 transcriptions.

Speaker diarization is the task responsible for identifying segments of an audio file by speaker, so that in each segment there is only one speaker. In our case, given a recording of an interview between a physician and a patient, we want to obtain an annotations file with segment timestamps (start and end) and corresponding speaker label, for the full duration of the recording, as exemplified in Table 4.1. We perform this task in a semi-automatic fashion. First, we automatically obtain segment timestamps determined to belong to different speakers. Then, manual segment validation is performed and all corrections applied, and, finally, the speaker is manually identified in each segment. This part is easily done by hand, since the interviewer voices are known beforehand and are expected to repeat the same questions in each recording. The only other voice is assumed to belong to the patient.

The first and automatic part of this task is performed in three steps. Given an audio file of a recorded interview, we first sample it at a 16kHz rate and extract 6 types of features per

Start	End	Label
0.0	14.500	Physician
14.500	31.820	Patient
31.820	32.400	Noise

Table 4.1: Example of an annotation file with speaker segments identified and labeled.

frame, which is defined as a window of 20ms of audio signal. These features are the root-mean-square (RMS), the spectral centroid, bandwidth, contrast, and flatness, and the zero-crossing rate, as defined by the librosa library (McFee et al., 2015). Secondly, we estimate a Gaussian Mixture model (Reynolds, 2009) with 3 components over the data points on the feature space. The choice of the number of components is based on the expected structure of each recording: scripted turns of dialog between the physician and the patient, rarely with interruptions, and an ever-present background noise, which may sometimes superimpose the dialog. Thus, we assume that the data points generated by each source fit a normal distribution (for the physician, patient, and noise) and that each is distinct enough to be differentiated. Finally, with this estimated model over the feature data, we assign each frame to a cluster, as given by the component in the mixture model with the highest posterior probability for that sample. We apply a filter to the cluster assignments so that we observe a sequence of cluster blocks that more accurately represents scripted dialog turns, as exemplified in Figure 4.3. This filter is a simple pass-through over the frames with a window of 100 frames, where all frames in that window are assigned the respective window mode.

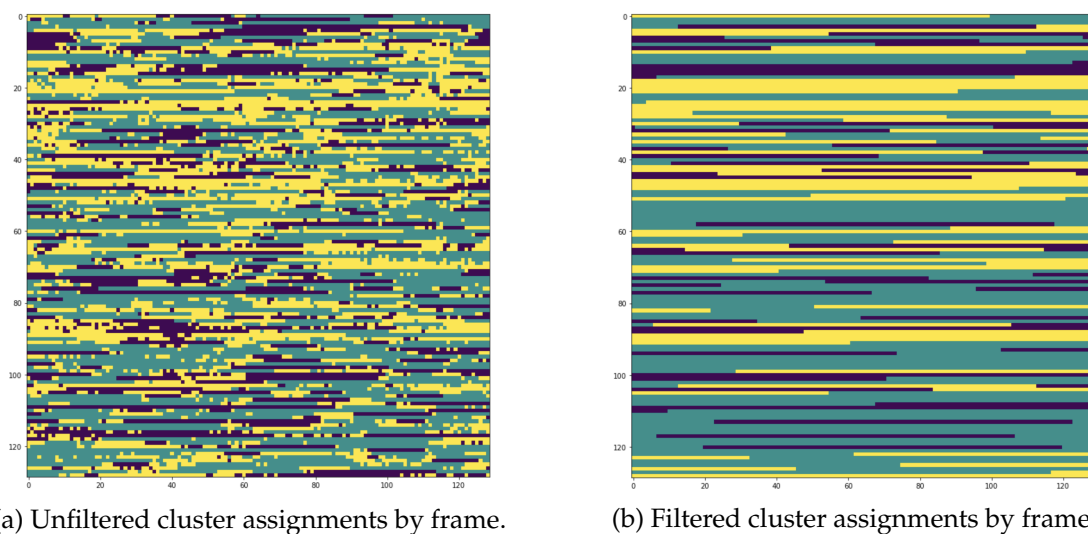


Figure 4.3: Cluster assignments by frame, represented in a square matrix, as given by a Gaussian Mixture model over the feature data.

The preparation pipeline facilitates processing in two ways, (i) it allows for the study of the patient’s verbal description separate from the interviewer’s speech, and (ii) since the dialog turns follow a specific script, the fragmentation is done automatically, separating each audio file into the different questions and answers, so that they can also be processed independently.

### 4.3 Baseline dataset and challenges

The baseline dataset is comprised of 94 rows (patients) and 29 columns (inputs). Table 4.2 summarizes the available inputs for each patient.

Parameter	Meaning and values	Other labels
ID	Unique identifier per patient	
Age	Age of the patient.	Idade (anos)
Gender	Gender of the patient: [Female, Male].	Género
Education (level)	Highest level of education obtained: [Basic, High School, Bachelor, Master].	Educação (nível)
Education (year)	Last completed year of the corresponding education level.	Educação (ano)
Occupation	Current professional occupation when applicable, otherwise refers to the latest professional occupation.	Ocupação
Activity status	Whether the patient is professionally active or not: [Active, Medical leave / Retired].	Activo
Pathology	Associated pathology: [Rheumatoid Arthritis, Spondyloarthritis, Osteoarthritis, Psoriatic arthritis].	Patologia
VAS pain	Pain intensity as reported by the patient, where 0 is no pain and 100 is the maximum pain ever felt by that patient: [0, 100].	EVA dor %
VAS disease	Disease state as reported by the patient, where 0 is very well and 100 is as worse as it has ever gotten for that patient: [0, 100].	EVA doença %
Evolution (years)	Number of years the patient has reported health problems associated with the disease (including pain).	Evolução (anos)
Duration (years)	Number of years that the patient has been diagnosed with the associated pathology.	Dur. (anos)
Therapeutics	List of applied therapeutics.	
ESR	Analytical parameter of the activity of the disease (Erythrocyte Sedimentation Rate).	VS
RCP	Analytical parameter of the activity of the disease (Repetitive C Protein).	PCR
Q1-Q7	Document of text corresponding to the labeled question of the interview (total of 7 questions).	
A1-A7	Audio file corresponding to the labeled question of the interview (total of 7 questions).	

Table 4.2: Dataset inputs and corresponding description.

The nature of the data used in this study presents a set of challenges to the task of modeling it in terms of its semantic and syntactic structures. Three types of challenges were found in the data, relating to the background and characteristics of the interviewed patient, the quality of the audio and textual data, and, finally, the availability. All of these challenges condition the

applicability of any type of analysis, linguistic or paralinguistic.

Firstly, we are concerned with the content and nature of the data, which is linked to the variety of ages, backgrounds, and personalities of the subjects included in the study. Specifically, the complexity and detail of the discussed topics ranges from vague to dense, simply because some people naturally speak more than others, resulting in denser documents, whilst others restrict themselves to short, precise or imprecise, answers. Additionally, the relationship established between the physician and the patient also restricts, or elicits, the development of the thought process. These characteristics render a collection of semantically related documents, although of different lengths, vocabularies, development, and precision.

Secondly, we are concerned with the quality of the obtained data. Since the textual documents are the result of transcribed speech, they inherit some speech disfluencies which could not be mitigated with a clean transcription strategy, such as the lack of syntactic coherence, which sometimes results in incoherent phrases. Regarding the audio quality, because the recordings were captured in the medical office without professional equipment (in an attempt not to intimidate the patients), the automatic processing is very limited.

Finally, we are concerned with the amount of available data to perform the analysis. If the patient's answers to all interview questions are concatenated into a single document, there are a total of 94 long documents, which is a very limited amount for almost all types of analysis, resulting in statistically irrelevant conclusions. If the fragments are considered independent, we would have  $94 \times 7$  documents, albeit short-text. The resulting conclusions could be statistically sounder, but the information is also harder to extract, due to their short length, and the fact that the notion of a patient could be lost.

## 4.4 *Summary*

In this chapter we presented the dataset used in the study, including the collected information and the preparation pipeline, which includes semi-automatic diarization of the audio, fragmentation into question segments, and manual transcription. Additionally, we discussed the challenges associated with the nature of the data, specifically the intrinsic variability of the content of the descriptions of pain, the syntactic incoherence of the text, the poor audio quality, and the limited availability of data.



# 5 Characterization of the Population

In the present chapter we aim at characterizing the population of patients experiencing symptoms of chronic pain in a space of linguistic features, as determined by their natural language descriptions of the experience. We define this characterization as both the mapping of the population onto the feature space, and the definition and quantification of any relations found in that space, as given by intrinsic qualities or extrinsic parameters.

Given the baseline dataset presented in the previous chapter, this experiment is performed in two main steps. First, the projection of the population on the linguistic feature space. Specifically, these features are based on topic modeling techniques, so that each patient is mapped onto a latent semantic space representing the aspects discussed in the collection of descriptions. This is the method used because it allows us to identify topics and quantify their importance for each patient in an unsupervised manner, as determined by the scripted interviews used to generate the descriptions, which guide the patients to reflect on the cognitive aspects determined by the literature as the most important for pain assessment and management. The second and last step encompasses the analysis of the projected descriptions. This includes similarity measures between distinct patients, clustering analysis, and semantic characterization of these groups and the ones defined by objective demographic and clinical parameters.

The structure of this chapter is as follows. First, we lay out the experimental setup of both topic modeling and characterization of the population. This includes the definition of the analysis and evaluation metrics. This is followed by a results and discussion section.

## 5.1 *Topic modeling*

We are interested in obtaining a projection of the patients on a latent semantic space. Specifically, a matrix projection  $T$  of  $n$  patients on the topic space ( $n \times k$ ), and the corresponding distributions of weights over the vocabulary for each topic, for  $k$  topics, unknown beforehand.

Our approach is based on the fragmented documents (7 documents per patient). We have decided on this approach because, otherwise, we would be restricted to a collection of  $n = 94$  documents. The fragmented approach means that, for the purpose of topic modeling, we are considering each fragment as an independent document, and consequently, with an independent projection. Matrix  $T$  is obtained by aggregating the projected fragments by patient. We perform this aggregation by averaging each topic importance over the corresponding 7 fragments, which assumes that all fragments (answers to each question in the interview) have equal importance for the description of the experience of that patient. We start by preprocessing the text and defining the topic models to apply, and, finally, define the evaluation process used to determine the most adequate fragment projection on the topic space. This will be the latent space used to characterize the population, after fragment aggregation by patient.

### 5.1.1 Text preprocessing

This task is in charge of noise removal and standardization of text. The applied techniques are, sequentially, text lemmatization (which includes identification of collocations and Part-Of-Speech (POS) tagging), and stop-word removal. The following describes each of these techniques, as well as the methods and tools used to apply them.

Lemmatization is the process in charge of obtaining the root word (lemma) of any given word in a sentence. The lemma is necessarily a valid word with semantic value. The objective is, thus, to standardize the text, since the variability intrinsic to text, such as conjugations of the same verb, variations of the same noun, and others, are all reduced to the same lemma. The tool used for this process is STRING (Mamede, Baptista, Diniz, & Cabarrão, 2012), which also tags each lemma with the original word's POS tag and finds collocations, which are sequences of two or more words which are commonly used together to define a specific concept. Collocations are defined as a single token, given that considering each of the composing words independently does not render the same semantics. Some examples are the n-grams "*senhor doutor*" (health professional) and "*mais ou menos*" (more or less). After lemmatization the vocabulary size drops 27%, which indicates a significant level of syntactic variability in the original text.

Stop-word removal excludes from the lemmatized and POS tagged text words such as determinants, pronouns, auxiliary verbs, conjunctions, and prepositions, which by nature do not convey significant semantic information relevant outside the syntactic context of the phrase,

thus producing only noise when under topic analysis. As a baseline set of stop-words, we use that defined by the NLTK package (Loper & Bird, 2002), for the Portuguese language, together with a filter for any POS tag that is not a verb, noun, adjective, or adverb. Additionally, we also remove any tokens that have a document frequency lower than 2 documents, or higher than 95% of the documents in the collection. This is so because certain tokens are so frequently used that they no longer carry information relevant to distinguish the documents, and, conversely, highly infrequent words (that are unique to one document) by definition distinguish too well each document, without taking into consideration the underlying semantics, and don't allow for an adequate generalization of the model. Stop-word removal yields a vocabulary 65% smaller, which highlights the lack of richness of the collection's vocabulary.

Table 5.1 reveals the top 20 words more frequent, before and after preprocessing, highlighting the importance of this initial step. This preprocessing pipeline yields a new version of the original documents, which is standardized, with noise removed, and with a total of 526 unique tokens. Figure 5.1 shows the probability distribution of this final vocabulary (ratio of number of documents in which each vocabulary word appears and the total number of documents in the collection). Expectedly, most words have a very low probability of occurring in the collection.

Before				After			
eu	de	muito	ir	andar	doer	querer	afetar
não	em	mas	assim	dia	sempre	tomar	melhorar
ser	estar	fazer	com	mão	medicação	mau	tempo
que	dor	saber	porque	poder	joelho	bastante	passar
ter	um	mais	mesmo	conseguir	pé	começar	esperar

Table 5.1: Top 20 words more frequent, before and after preprocessing.

### 5.1.2 Models

The presented NMF and LDA models, which, as discussed, are expected to have a limited performance in the setting of short-text documents, are applied as baselines. The described SeaNMF and CluWords models have been shown to have the best performance in a similar setting to the one described in the present experimental setup, and, thus, are applied to further explore the data and overcome its challenges. We apply both these models due to their varying nature, since SeaNMF does not resort to external information but is limited by the collection's size, and CluWords resorts to external information but is limited by domain adaptability and

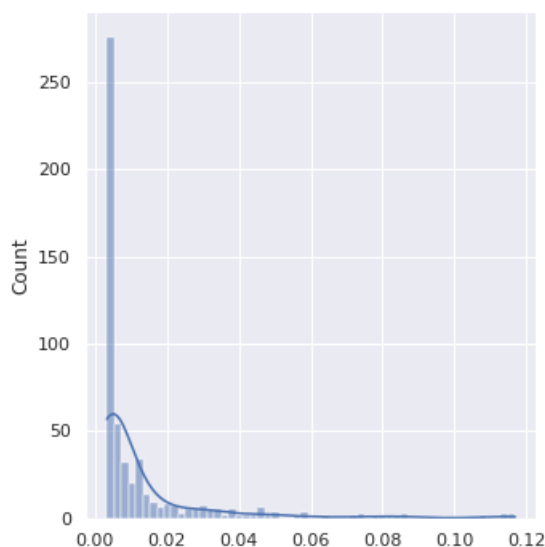


Figure 5.1: Distribution of vocabulary probabilities.

poor vocabulary. We explore these domain adaptability concerns when using external word-embedding models, specifically by comparing the performance of CluWords with different word-embedding models, specifically FastText and BERT (Devlin, Chang, Lee, & Toutanova, 2018), which have been pre-trained on Portuguese corpora. These are summarized in Table 5.2.

LDA
NMF
SeaNMF
CluWords (FastText)
CluWords (BERT)

Table 5.2: Considered topic models for the experimental results.

### 5.1.3 Evaluation

Given that this is an unsupervised task, the evaluation that we can performed is solely based on intrinsic qualities of the modeling of the collection in the topic space. There are two main types of intrinsic evaluation. First, interpretability metrics, which are concerned with the semantics associated with the projection and the relation with the nature of the data under study, and, second, clustering metrics of the projected documents on the latent semantic space, which are concerned with evaluating the stowage of data points in the given space. These can be both context agnostic or context dependent. We evaluate the applied topic models under both of these types of evaluation.

### 5.1.3.1 Interpretability metrics

Our first concern is the number of topics to extract,  $k$ , which is unknown beforehand. Even though this is somewhat of an arbitrary choice, it may be informed by the nature of the collected data. Each question of the interview focuses on at least one aspect of the experience. It is up to the patient to develop, or not, the thought process to more aspects. Given this statement, the number of topics to extract should be bounded by a minimum of 7 topics (the number of questions in the interview). The value of  $k$  is determined based on empirical evaluation and human interpretation of the extracted topics.

Given a fixed number of topics to extract, following the literature, we evaluate the topic coherence of each topic model, given by the PMI score. Because we are dealing with an extremely low-resourced collection of documents, we focus on the Positive PMI (PPMI) metric, which adequately accounts for word pairs that never co-occur (which is a common problem in our collection). The PPMI metric is defined in Eq. (5.1), where  $t = 10$  is the number of top most weighted words of a topic.

$$\text{PPMI} = \frac{1}{t(t-1)} \sum_{i < j \leq t} \max\left\{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, 0\right\} \quad (5.1)$$

Finally, we evaluate topic modularity by determining the number of words shared between all topics extracted by one model (given the top  $t = 10$  most weighted words of each topic). A topic that does not share a single word with the remaining topics may define a concrete, modular concept, which allows for an independent evaluation of the projected population on that dimension. Therefore, we are most interested in the topic model which extracts topics with most unique words, however taking into consideration that the probability of choosing  $k$  sets of  $t = 10$  unique words from the vocabulary is not zero. Thus, the discussion around these results cannot be independent of the actual top words of each topic.

### 5.1.3.2 Clustering metrics

Regarding the clustering metrics in the modeling space, we start with the ones which are to a degree agnostic to the problem domain, in this case, how well the projected documents can be clustered in the latent space, and which is the most adequate number of clusters for the

samples. In this experimental setup, in which we are dealing with the fragmented short-text documents (around 20 tokens per document), the typical document projection is composed of a highly weighted dimension and the remaining with infinitesimal values. This is so because, most of the times, in such a short amount of words, semantically speaking, only one concept is being discussed. Given this characteristic, we expect to obtain the best clustering for a model with the number of clusters equal to the number of extracted topics. However, we do not expect to obtain a perfect clustering, as if all projected documents we restricted to a specific dimension by groups. Indeed, some documents may have higher weights on more than one topic. Thus, we look at the Silhouette Coefficient of each sample, defined by Eq. (5.2), in each topic model, for the number of clusters equal to the number of topics. This metric determines as a well-defined cluster that which has all points well-distanced from the next nearest cluster, and the mean distance between all points of that cluster is minimal, where  $a$  is the mean distance between a sample and all other samples in the same cluster, and  $b$  is the mean distance between that sample and all other samples in the next nearest cluster. This evaluation will give us insights as to how distributed are the samples across the clusters, and how adequate each sample is to the assigned cluster.

$$s = \frac{b - a}{\max(a, b)} \quad (5.2)$$

To conclude on the best model for our data, we also have to look at what is actually assigned to each cluster, and thus need to define a qualitative cluster evaluation, which is domain dependent. In our case, the interview design is such that each question should elicit specific thought processes on the answers, which, therefore, should be focused on specific semantical aspects. If the documents answering to a given question should share a similar, specific distribution of topics, then we should observe each cluster to have a distribution of documents over a very small number of questions (because different questions should resolve into different topic mixtures). In order to assess this statement, we define matrix  $M$ , of dimensions  $c \times q$ , where  $c$  is the number of clusters, and  $q$  is the number of questions (fixed at 7). Each entry  $M_{ij}$  is defined as the percentage of documents in cluster  $i$  which are answers to question  $j$ , such that each row  $M_i$  sums to 1. We also define a threshold value  $\beta \in [0, 1]$  such that if  $M_{ij} < \beta$ , then  $M_{ij} = 0$ . We use this value to discard residual values which do not provide statistically relevant information. We define as the most interpretable clustering model that which has the

most sparse matrix  $M$ , specifically, that which maximizes Eq. (5.3), for a given value of  $\beta$ .

$$\text{sparsity}(M, \beta) = \frac{\sum_i^c |\{j \in M_i : j = 0\}|}{c} \quad (5.3)$$

After this evaluation we obtain a topic space on which to characterize the patients. Because the topic space is obtained through the fragmented dataset, the notion of a patient topic projection may be recovered by aggregating the corresponding projected fragments.

## 5.2 Characterization

The previous sections defined how to extract and evaluate the semantic structures associated with each description of pain, by means of topic modeling. In this section, we define the methodology to visualize and discuss these structures on the latent semantic space, in order to compare patients, identify groups, and correlate with demographic and clinical parameters.

We perform this characterization following three approaches. First, we look at the projected population as a whole, and characterize it. This includes the interpretation and labeling of the extracted topics and topic importance mixtures, and the identification of the most common and important topics and words. Second, we split the population into groups of similar topic distributions, which represent the different types of experiences of pain, and characterize them independently. This encompasses all evaluations performed in the first step, and further correlation with demographic and clinical parameters, specifically regarding their distributions in these similarity-defined groups. This allows us to associate types of experiences of pain, according to their descriptions, to specific ranges or values of objective parameters. For the third step, we split the population into groups defined by the demographic and clinical parameters, and perform the previous analysis in these groups independently. This allows us to associate values or ranges of demographic and clinical parameters with aspects of experiences of pain.

The following sections describe in greater detail how each of these steps is carried out and which evaluations are performed. All steps assume a matrix projection  $T$  of  $f$  fragments on the topic space ( $f \times k$ ), as determined by one of the topic models in Table 5.2, and the corresponding distributions of weights over the vocabulary for each topic. Matrix  $T$  may be aggregated by various parameters, but, most importantly, by patient, obtaining a topic mixture for each.

### 5.2.1 Overall population

We start by interpreting the topic space by assigning a label to each topic, so that it is more concrete and easier to discuss. This is performed by interpreting the top 10 most weighted words of each topic. The topics may now be referenced by label and represent the whole concept associated with that label. Topic interpretation depends on context. These assigned labels are validated by aggregating matrix  $T$  by question ( $7 \times k$ ), which results in the mean mixture of topics by question, representing the average answer to each. Because each question in the interview focuses on a specific aspect of the experience of pain, this analysis also allows us to understand exactly how complex each aspect is to the patients, and which semantical topics it encompasses. Additionally, we can also determine the success of the interview in terms of its answers, and judge if the obtained descriptions of pain match the intention of the design.

The second analysis focuses on topic importance distribution. We want to determine how the importance of any given topic varies in the population, which have the highest and lowest importances, and which are discussed by more and less patients. To obtain this in terms of patients, we aggregate matrix  $T$  by patient ( $n \times k$ , where  $n = 94$  patients).

Finally, considering that each projected patient is defined by its top  $N \leq k$  most weighted topics, we define topic co-occurrence in a collection of projected patients the co-occurrence of these top  $N$  topics in that set. Topic co-occurrence dictates how correlated are the topics in a set of patients, and provides insights as to how the description of the experience of pain flows between the aspects or concepts that they represent.

### 5.2.2 Topic similarity clusters

This analysis is based on the premise that similar descriptions in the topic space relate to similar experiences of pain, such that each of these similar groups represents a type of experience. We also raise the hypothesis that these types of experiences are correlated with demographic and/or clinical parameters.

Given the projection matrix  $T$  aggregated by patient ( $n \times k$ ), we start by assigning cluster labels to each patient, so that we obtain groups of similar patient descriptions, based on topic mixtures. The number of clusters is determined based on the clustering model inertia value, the Calinski-Harabasz index, the Silhouette Coefficient, and Davies score, because there is no



obvious arrangement of the patients in groups, and these metrics provide insights on their stowage. Once the patients are arranged in groups, we perform an ad hoc interpretation of the obtained clusters by observing their mean topic mixture and variance, so that we may, first, validate their grouping, and second, interpret the different types of experiences of pain. To this end, we apply the same steps of analysis as for the whole population, but by cluster.

Finally, for each type of experience we analyze the distribution of demographic and clinical parameters, so that we can try to correlate semantic structures of descriptions of pain with values or ranges of objective parameters. This is not independent of the distribution of patients per parameter. In Tables A.1 and A.2 and Figure A.2 are presented these distributions (category values with a dash refer to patients that did not disclose such information).

### 5.2.3 Demographic and clinical clusters

In this final step of the characterization, we raise the hypothesis that certain characteristics of the patients under study influence their experiences of pain, specifically in terms of the mixture of topics of their descriptions.

We define groups of patients according to demographic and clinical parameters, and observe the descriptions of pain of these groups, in the topic space. The demographic parameters that we consider are sex, age, education level, and whether the patient is professionally active or not. The clinical parameters that we consider are pathology, pain intensity, duration of the disease, ESR, and RCP. We apply the same steps of analysis as for the whole population, but by group of patients as given by these parameters.

## 5.3 Results and discussion

In this section we present and discuss the results associated with the evaluation of the experimental setup of the characterization of the population. The goal is to decide on the best topic modeling of the fragmented data, aggregate it by patient, and use it for characterization. We start by defining the model parameters and then discuss the evaluation results, for both topic modeling and characterization.

### 5.3.1 Topic modeling

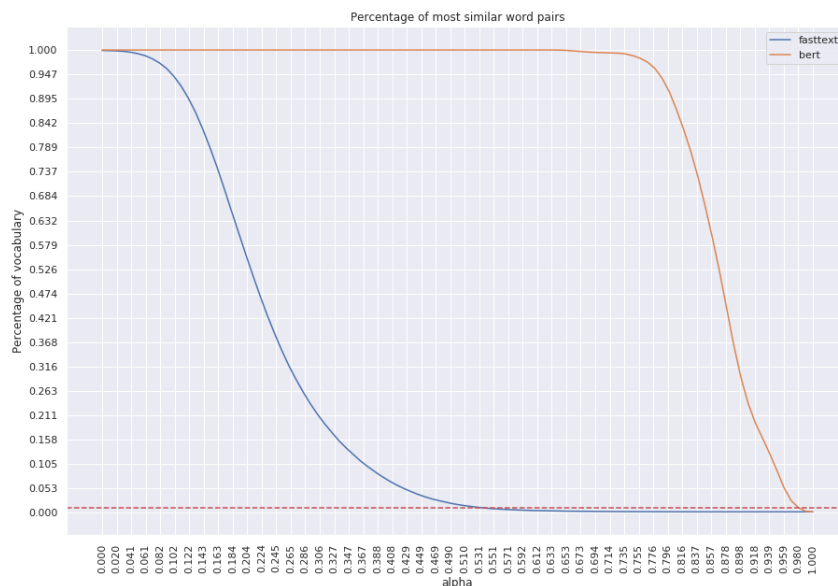


Figure 5.2: Mean percentage of tokens in the vocabulary considered sufficiently similar to any other token as the parameter  $\alpha$  increases. The horizontal dashed line represents the 1% mark, which, for our limited vocabulary, translates to roughly 5 words.

The LDA model parameters are those defined as default in the Sci-Kit Learn package (Pedregosa et al., 2011). The NMF model, also from the same package, is set to initialize the procedure with random factoring matrices, and use the coordinate descent solver. The SeaNMF model implementation is the one provided by the authors and the parameters are those defined as default. Finally, since there was no public implementation of the CluWords model provided by the authors, at the time of writing, it was implemented specifically for this study. As described in the original paper, when the CluWords  $\alpha$  parameter is set to 0, the whole vocabulary is considered sufficiently similar to any other term, softening the TF-IDF (CluWords) counts to the maximum and possibly diluting important information in the whole vocabulary. Conversely, when  $\alpha = 1$ , only a single token is considered sufficiently similar (the token itself), replicating the behavior of the original TF-IDF method. However, the rate of drop in percentage of similar terms, as  $\alpha$  grows from 0 to 1, is dependent on the richness of the vocabulary and the meaning of the word-embeddings derived from the external model. This rate can be seen in Figure 5.2, for our vocabulary, regarding the FastText and the BERT word-embedding models. Observing this figure, we conclude that, for the BERT word-embedding model, 100% of our vocabulary scores more than 0.65 on the cosine similarity distance metric (vertical line,

where  $\alpha = 0.65$ ). This can be explained by the fact that the BERT model already incorporates contextual information, restricting our vocabulary terms to a specific region in space, which in turn translates to very similar vectors. This is also evidenced by the steep drop observed when  $\alpha$  gets closer to 1. On the other hand, for the FastText model, even though we also observe a steep drop, it occurs much behind and more smoothly, hinting that the vocabulary is better distributed on this word-embedding space. Table 5.3 reveals the top 5 words with higher cosine similarity scores for two random samples of tokens in the vocabulary, and their corresponding score. For the following experiences, we fix  $\alpha = 0.55$  for the FastText model, and  $\alpha = 0.98$  for the BERT model, so that, on average, the TF-IDF (CluWords) counts are smoothed over 1% of the vocabulary, with either model (which, for our vocabulary, roughly translates to 5 words).

FastText		BERT	
cansaço	aguentar	cansaço	aguentar
desconforto (0.65)	cansar (0.61)	cansar (0.95)	segurar (0.94)
cansar (0.63)	suportar (0.59)	cansado (0.95)	avançar (0.94)
sensação (0.59)	incomodar (0.57)	desconforto (0.95)	doer (0.94)
cansado (0.58)	esperar (0.57)	ardor (0.94)	apanhar (0.93)
inchaço (0.58)	parar (0.57)	inchaço (0.93)	moer (0.92)

Table 5.3: Top 5 words with higher cosine similarity scores for a random sample of tokens in the vocabulary, and their corresponding score. Note that with the proposed thresholds of  $\alpha$  only a select few of these words would be used.

### 5.3.1.1 Interpretability

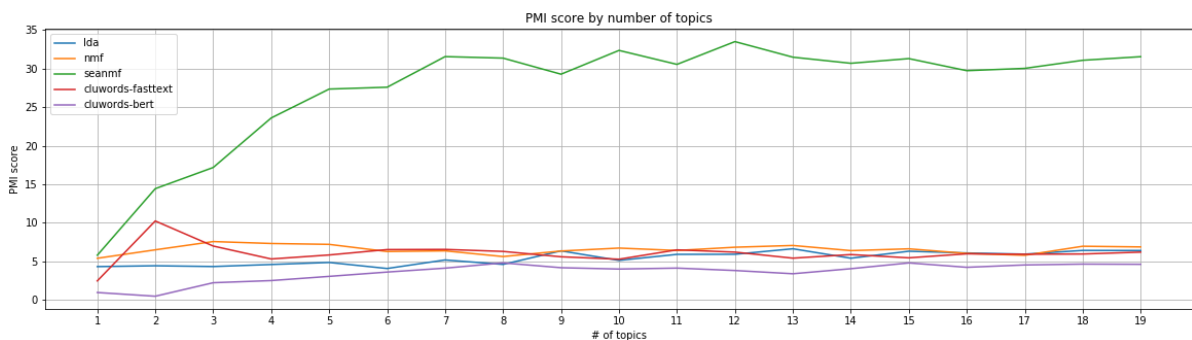


Figure 5.3: PPMI score of each model, over a range of topics, on the dataset vocabulary.

The PPMI score assigns a higher score to topic models which extract more coherent topics, where coherence is defined as most weighted words that most commonly co-occur in the collection of documents. Figure 5.3 plots the scores for all models, across a wide range of

Topic	Top 10 words
0	braço, perna, trabalho, joelho, andar, casa, difícil, posição, conseguir, anca
1	mau, dia, poder, sempre, manter, afetar, começar, resto, hora, noite
2	começar, movimento, andar, articulação, rigidez, hora, corpo, constante, provocar, menos
3	melhorar, tratamento, doença, tomar, medicação, tempo, passar, grande, querer, conseguir
4	pensar, esperar, evoluir, ver, origem, medicação, ideia, andar, problema, ano
5	andar, joelho, sempre, pouco, costa, profissional, emocional, pé, trabalho, dia
6	mal, costa, doer, tempo, chegar, profissional, altura, começar, pescoço, conseguir
7	dia, medicação, tomar, mau, doer, sempre, sensação, piorar, continuar, andar
8	afetar, sempre, poder, doença, dia, normal, andar, querer, gente, entender
9	levantar, poder, explicar, artrite reumatóide, mexer, conseguir, origem, pé, dia, noite
10	querer, conseguir, poder, mexer, andar, mão, gente, melhorar, braço, parar
11	mão, joelho, pé, doer, anca, articulação, bastante, ombro, pulso, osso

(a) LDA

Topic	Top 10 words
0	doer, dia, mal, depender, começar, igual, esforço, conseguir, noite, passar
1	origem, doença, osso, ideia, artrite reumatóide, problema, reumatismo, perceber, explicar, mínimo
2	bastante, lepicortinolo, vida, profissional, de vez em quando, trabalhar, levar, chegar, altura, relação
3	pensar, costa, continuar, problema, artrite, pouco, evoluir, igual, sei, nível
4	afetar, poder, trabalho, casa, trabalhar, profissional, normal, amanhã, físico, explicar
5	medicação, tomar, sempre, continuar, tempo, deixar, melhora, n vezes, desaparecer, suspender
6	mão, articulação, ombro, joelho, fechar, problema, conseguir, pé, mal, força
7	querer, conseguir, gente, mexer, braço, corpo, sempre, emocional, para trás, parar
8	andar, mau, estar, ver, perna, custar, piorar, trabalho, sempre, conseguir
9	melhorar, tratamento, piorar, grande, começar, medicamento, resultar, deus, ver, verdade
10	esperar, evoluir, manter, continuar, mau, ideia, tempo, estável, ver, depender
11	joelho, pé, anca, ombro, pulso, cotovelo, articulação, corpo, dedo, costa

(b) NMF

Topic	Top 10 words
0	pessoal, tentar, ultrapassar, profissional, nível, aspeto, pessoa, causar, influenciar, resto
1	mão, pé, joelho, doer, braço, pescoço, perna, articulação, parecer, fechar
2	leve, lidar, pedir, tarefa, executar, pesado, colega, trabalho, sair, gerir
3	possível, forma, nada, limitado, de lado, saúde, estável, gostar, manter, esperar
4	chuva, mudança, descrever, nevoeiro, partir, na totalidade, saída, nervoso, apanhar, desculpar
5	presumir, muscular, análise, especial, intermédio, conversa, melhora, surto, praticamente, comum
6	palavra, fixar, frente, temporada, vacina, cheio, subir, escada, crise, acabar
7	impossível, janeiro, desconforto, recorrer, aos bocados, cortar, supermercado, tocar, picar, secretário
8	impressionante, artrite, arrastar, momentâneo, artrite reumatóide, quente, incapacitante, fraco, tempo quente, ponto
9	assinar, repetitivo, intensidade, ontem, aumentar, a seguir, diferença, demorar, almoço, cortisona
10	esposo, metro, de um lado para o outro, diminuir, estacionamento, fundo, carro, sofrer, a pé, caminhada
11	valongo, hospital, conta, encontrar, horrível, médico, mandar, boca, ui, cama

(c) SeaNMF

Topic	Top 10 words
0	começar, parar, esperar, voltar, demorar, acontecer, continuar, acabar, sair, aguentar
1	medicamento, tratamento, medicação, metotrexato, fisioterapia, reumático, pomada, reumatismo, cortisona, tomar
2	perna, ombro, joelho, dedo, cotovelo, tornozelo, pescoço, tendão, mão, punho
3	conseguir, pegar, tirar, tentar, chegar, voltar, encontrar, falhar, perder, ajudar
4	afetar, causar, provocar, depender, resultar, influenciar, alterar, controlar, diminuir, aumentar
5	artrite, doença, artrite reumatóide, inflamação, pericardite, reumatismo, infeção, reumático, medicação, inflamatório
6	de um lado para o outro, de vez em quando, de um momento para o outro, de cada vez, para sempre, para trás, trabalho de casa, dia de amanhã, ter a ver, de repente
7	bastante, menos, pouco, mínimo, mau, quase, praticamente, totalmente, ideia, mal
8	querer, chatear, cansar, apetecer, pensar, esquecer, esforçar, incomodar, gostar, tentar
9	entender, perceber, explicar, perguntar, presumir, responder, pensar, desculpar, falar, considerar
10	melhorar, melhora, melhoria, diminuir, aumentar, ajudar, alterar, esforçar, piorar, agravar
11	osso, músculo, ilíaco, ósseo, pescoço, cervical, costa, lombar, origem, muscular

(d) CluWords (FastText)

Topic	Top 10 words
0	perceber, conseguir, levar, acontecer, aparecer, levantar, nomeadamente, considerar, envolver, influenciar
1	enorme, comum, igual, cheio, pegar, adaptar, bastante, julgar, junto, entretanto
2	osso, perna, peito, dedo, correr, cheio, braço, doer, andar, prender
3	querer, ligeiramente, ultrapassar, cair, suportar, sar, frequente, controlar, pedir, parar
4	mau, mal, dia, andar, costa, estar, querer, para trás, menos, tentar
5	medicação, tomar, dia, sempre, continuar, doer, andar, poder, começar, normal
6	joelho, ombro, pé, anca, cotovelo, pulso, punho, apanhar, constante, alto
7	esperar, evoluir, continuar, manter, aumentar, tentar, pensar, ver, estável, desejo
8	forte, intenso, leve, umar, muitas, alguma, tenhar, raro, ligeiramente, essar
9	mão, articulação, doer, pé, costa, problema, fechar, principal, horrível, punho
10	melhorar, melhora, alterar, mudar, tratamento, piorar, grande, para já, deus, comar
11	origem, doença, ideia, artrite reumatóide, pensar, costa, mínimo, reumatismo, problema, anca

(e) CluWords (BERT)

Table 5.4: Extracted topics by each model with hand-assigned labels.

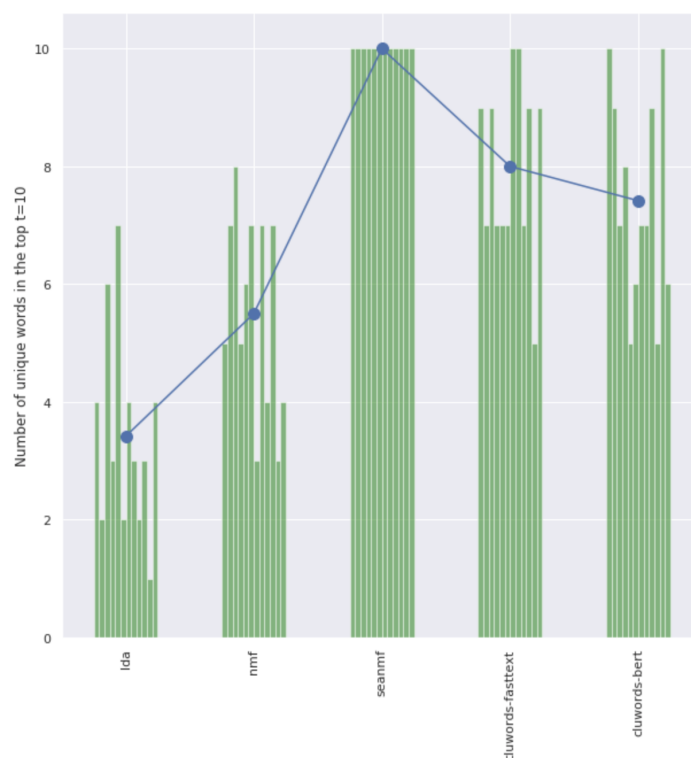


Figure 5.4: Modularity of each model’s extracted topic, presented in tables 5.4. Defined as the number of words in a topic (given its top  $t = 10$  most weighted) that are unique to that topic for all other topics extracted by the model. Mean model modularity is also shown in blue.

topics, which also allows us to validate our choice of the number of extracted topics. We observe a clear distinction between SeaNMF and all other models. This model looks at word co-occurrence in the collection and extracts word and context vectors for each term in the vocabulary. In practical terms, it promotes to belonging to the same topic terms that frequently share similar contexts, even if they never co-occur in the collection. This type of contextual information is designed to overcome the limitations of short-text documents, specifically those carried to the baseline BoW or TF-IDF representations. Although there is a limited amount of samples, the extracted contextual vectors seem to allow for a superior topic coherence. On the other hand, CluWords, with either FastText or BERT, does not seem to outperform the baseline LDA and NMF models, as suggested by the literature. This limitation can be attributed to domain adaptability concerns, which are highlighted in our context by the highly contextual meaning of the words employed by the patients when describing a personal experience, often resorting to linguistic tools such as analogies or metaphors, and the poor variety of the vocabulary. If synonyms or words describing similar concepts are not employed, the TF-IDF smoothing done by CluWords is rendered practically ineffective. We also observe a marginally

lower score for CluWords with contextual word-embeddings provided by BERT. This is expected, following the previous discussion on the cosine similarity of the vocabulary on both of the word-embedding spaces.

Regarding the number of topics to extract, we focus on values higher than 7 (the number of questions in the interview). Analyzing the interview structure, and having in mind the actual answers given by patients, we conclude that a few questions did not elicit the development of other cognitive aspects other than the question itself (e.g. questions 1, 5, and 7). The remaining questions have varied thought elaborations. Given these statements, we decide on fixing the extraction to  $k = 12$  topics, and should thus be considered the baseline from hereon. Observing the top  $t = 10$  words for each extracted topic by each model in Table 5.4, we make the following observations. Starting with the baseline models, we conclude that the top words defining the NMF topics allow for a slightly easier interpretation than those of LDA, even though their corresponding PPMI scores are practically identical (indeed, NMF is marginally superior). Nevertheless, both topic models are still hard to interpret. Observing CluWords (FastText) topics, it becomes dramatically easier to interpret. Indeed, some seem to relate to concrete concepts, such as pain location, intensity, and treatment. However, again, this model is indistinguishable from the baselines and CluWords (BERT), according to the PPMI score. SeaNMF, on the other hand, which has the greatest coherence score, seems to be extremely overfit to the collection, with very hard to interpret topics. In Figure 5.4 is shown the modularity of each model. In this case, as previously defined, topic modularity is given by the number of words in the set of most weighted words of a topic that are unique to that topic (for a model extracting  $k$  topics). This metric corroborates the topic interpretation that was previously discussed, because a model with more modular topics may have more well-defined concepts represented by the topics.

These observations, together with the distribution of vocabulary probabilities in Figure 5.1, allow us to determine that a probability-based evaluation of topic coherence is inadequate for our collection. First, the number of samples, even though extended through fragmentation, is very limited, and, second, the vocabulary is extremely poor, with most words having a very low probability of occurring in the collection. Additionally, we conclude that SeaNMF is capable of having higher PPMI scores simply by selecting for each topic words that commonly share the same context (in this case, the context window is each document), producing semantically inferior topics.

## 5.3.1.2 Clustering

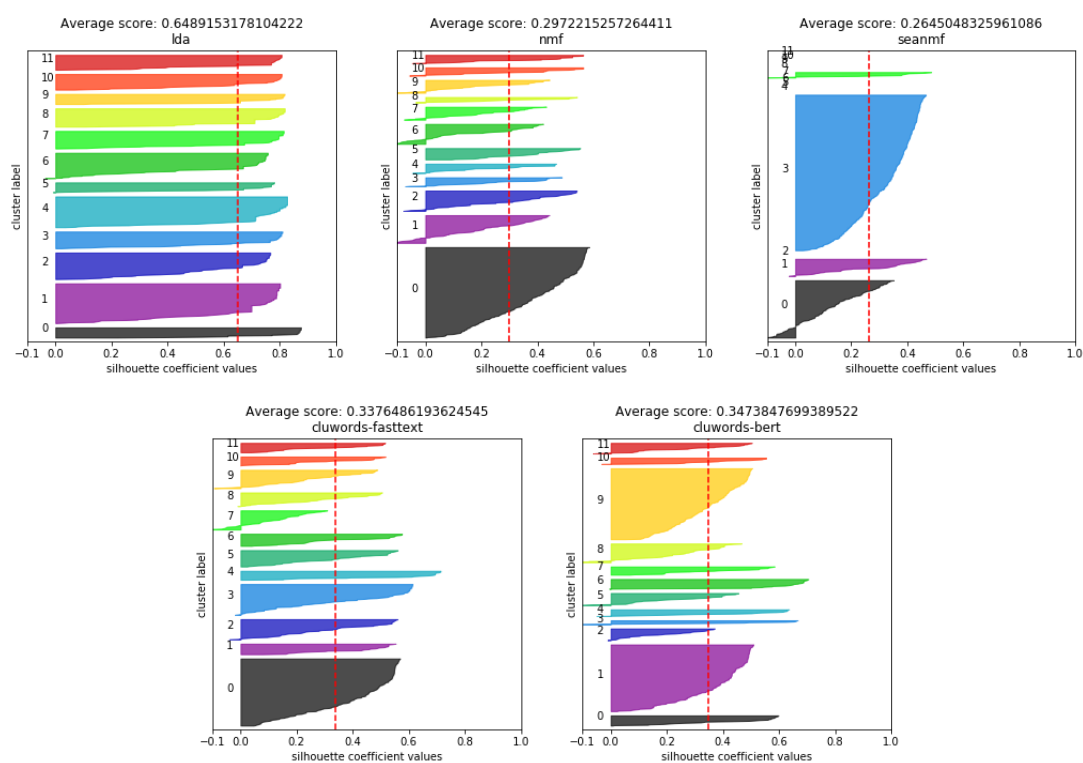


Figure 5.5: Silhouette of each model, for 12 clusters (equal to the number of topics). Represents the silhouette score assigned to each sample.

Fixing the number of clusters to equal the number of topics, we can observe the concrete silhouettes of each model in Figure 5.5. Observing the silhouette rather than simply its mean value gives us a clear understanding of the distribution of samples in space and in relation to the assigned clusters, so that we can better assess how well the projected documents are grouped in the topic space. The silhouette of the LDA model represents the ideal silhouette of a quality clustering of samples in a given space: the samples are evenly distributed across the clusters, there are no significant differences in the silhouette scores of a cluster's samples, all scores are well above zero (meaning cluster samples are well separated from other clusters), and there are no negative values (which imply wrong cluster assignments of the corresponding samples). Due to the statistical inference nature of LDA, the lack of instances (documents), and their short length nature, the documents are practically projected onto single dimensions on the LDA topic space, which results in an almost perfect clustering. All other models have a far worse silhouette for this number of clusters and topics. Even though the SeaNMF model has the highest topic coherence score, its silhouette indicates that the majority of the documents

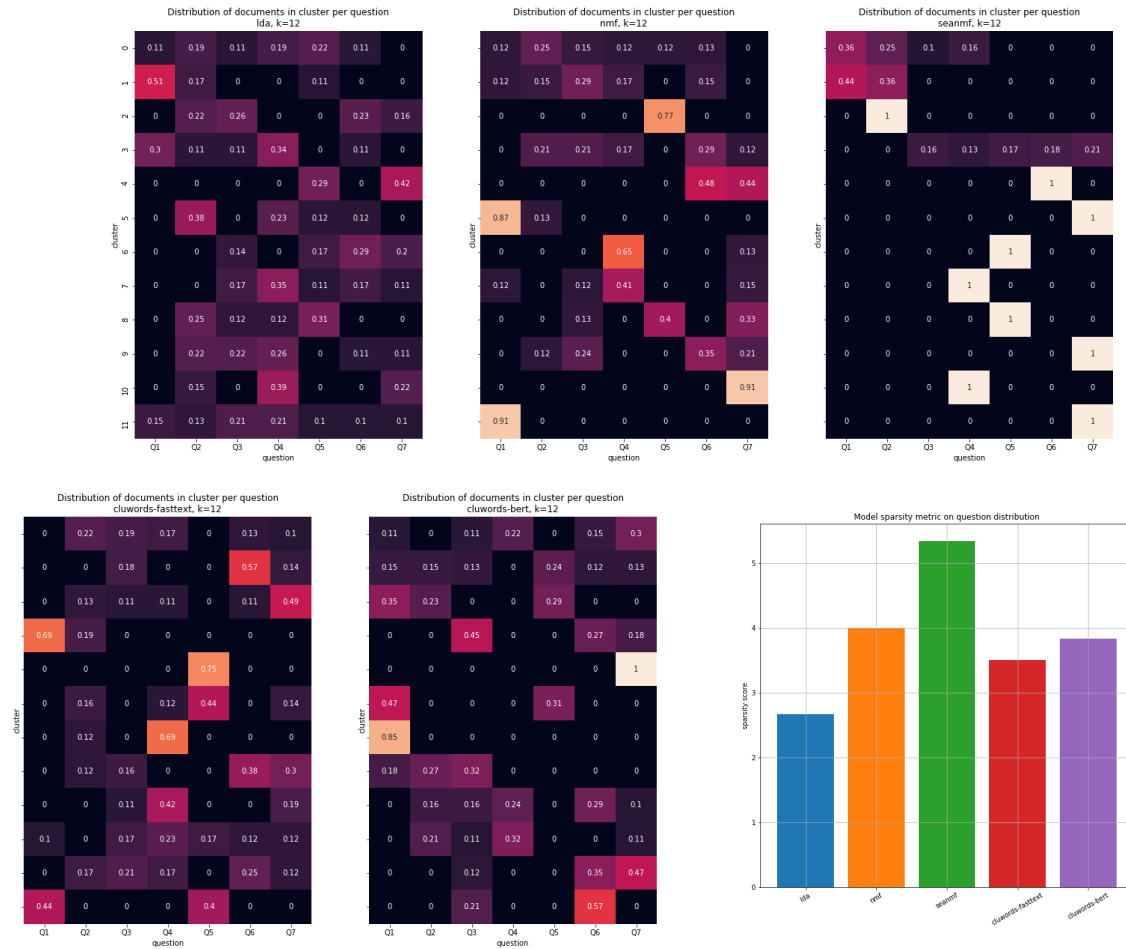


Figure 5.6: Matrix  $M$  for each topic model, of dimensions  $c \times q$ , where  $c = 12$  is the number of clusters, and  $q = 7$  is the number of questions. Threshold value  $\beta = 0.1$ . Inferior right corner is the sparsity score of the clustering over the topic space of each model, as given by Eq. (5.3).

are put into the same cluster (and, indeed, some of these have scores close to zero), or are poorly assigned to poorly-defined clusters (specifically observing cluster 0, with almost half of negative scores). For the remaining models, both CluWords models have higher mean scores than the baseline NMF. After all, CluWords builds on top of the TF-IDF representation, relying on the same NMF model parameters and implementation to factorize the representation matrix, albeit slightly more informative.

Finally, the sparsity score, specifically designed for our dataset and data collection protocol, allows us to get a high-level understanding of the contents of the created clusters on the topic space, in relation to the design of the interview. The results for each model can be observed in Figure 5.6, with the corresponding distribution matrix. This score is designed to express how well discriminated are the documents regarding their topics and question distribution.



Naturally, since LDA is a generative model, rather than discriminative, its score is expected and observed to be inferior than factorization-based approaches. By smoothing the TF-IDF counts with semantical information from external word-embedding spaces, CluWords is also expected to have a poor discrimination, albeit with more coherent topics than LDA. On the other hand, NMF is designed to discriminate, and that can be observed by the increased score when compared with the previous 3 models. Finally, the SeaNMF model exhibits the highest sparsity score, however, connecting with its silhouette in Figure 5.5, we quickly conclude that it is an artifact due to the poor distribution of samples per cluster (which the designed sparsity metric does not account for).

According to the previous observations and discussion, we discard the LDA topic space, because the extracted topics are hard to interpret, their corresponding most weighted words are highly shared among them, and we conclude that the almost perfect clustering of fragments in the topic space has the least relation to the interview scheme, which suggests that the obtained topic mixtures are less meaningful in this context than the remaining. We discard the SeaNMF topic space, as it is shown to be considerably overfit, with apparently meaningless topics, according to their most weighted words. The remaining models are all based on the same NMF model implementation and parameters, albeit on top of slightly different vocabulary-based representations of the fragment collection. Based on the observed results, we decide that the topic space given by CluWords (FastText) should be used to further characterize the population.

### 5.3.2 Characterization

In this section we present and discuss the results associated with the characterization of the population on the latent semantic space, obtained via topic modeling, as defined by the results and discussion of the previous section. In this case, it is the one extracted by CluWords (FastText), with  $k = 12$  topics, presented before, repeated here with additional labels in Table 5.5.

A topic is a distribution of weights over the vocabulary employed by the patients. The top most weighted words of each topic are the ones most commonly used in similar contexts by the patients, and, thus, are expected to relate to some concept being discussed in said contexts. This concept is not necessarily concrete nor comprehensive, and may be context dependent. Therefore, assigning a label to a topic is both a difficult and a somewhat biased task (this may be attenuated by having a third party perform the task, without providing information regarding

Topic	Top 10 words	Label
0	começar, parar, esperar, voltar, demorar, acontecer, continuar, acabar, sair, aguentar	Activity
1	medicamento, tratamento, medicação, metotrexato, fisioterapia, reumático, pomada, reumatismo, cortisona, tomar	Treatment
2	perna, ombro, joelho, dedo, cotovelo, tornozelo, pescoço, tendão, mão, punho	Specific locations
3	conseguir, pegar, tirar, tentar, chegar, voltar, encontrar, falhar, perder, ajudar	Actions
4	afetar, causar, provocar, depender, resultar, influenciar, alterar, controlar, diminuir, aumentar	Impacts (1)
5	artrite, doença, artrite reumatóide, inflamação, pericardite, reumatismo, infecção, reumático, medicação, inflamatório	Causes
6	de um lado para o outro, de vez em quando, de um momento para o outro, de cada vez, para sempre, para trás, trabalho de casa, dia de amanhã, ter a ver, de repente	Time intervals
7	bastante, menos, pouco, mínimo, mau, quase, praticamente, totalmente, ideia, mal	Intensity
8	querer, chatear, cansar, apetecer, pensar, esquecer, esforçar, incomodar, gostar, tentar	Impacts (2)
9	entender, perceber, explicar, perguntar, presumir, responder, pensar, desculpar, falar, considerar	Reflections
10	melhorar, melhora, melhoria, diminuir, aumentar, ajudar, alterar, esforçar, piorar, agravar	Evolution
11	osso, músculo, ilíaco, ósseo, pescoço, cervical, costa, lombar, origem, muscular	Generic locations

Table 5.5: CluWords(FastText) with topic labels.

the source of the data), which can introduce errors and questionable expectations. Having this in mind, and because each question in the interview aims at specific aspects of the experience of pain, using only the top 10 most weighted words of each topic, we interpret and associate a label reminiscent of the questions in the interview script. Some topics are more concrete and easier to interpret: treatment, specific locations, impacts (1, 2), time intervals, evolution, and generic locations. Others were associated with the aspects in the interview that are most related: activity, actions, causes, intensity, and reflections.

There are two important remarks regarding the assigned labels. First, each label is associated with an idea or concept that is more embracing than the top 10 words that suggested it in the first place. Because, from hereon, topics will be referenced by label, the top words should be referenced when making any statements related to the underlying semantics of descriptions of experiences of pain. Second, regarding the topics that apparently relate to the same idea, specific and generic locations, and impacts (1) and (2): the fact that these were extracted into separate topics tells us that their corresponding words were commonly used in different contexts, either because the semantical structures used to reference each sub-concept are different (e.g. specific versus generic locations of pain may be referenced differently due to

their specificity nature), or because they relate to actually different concepts and were poorly interpreted. This matter may be assessed by understanding the contexts in which each topic, or sub-concept, is used.

### 5.3.2.1 Overall population

Each question in the interview aims at specific aspects of the experience of pain, which are self-explanatory. By aggregating the topic importance by question, we can both understand in which context each topic is being used, and attempt to explain each aspect of the experience of pain not by its theoretical attributes, but rather by the observed mixture of topics (assuming the patient answers are related to the questions). The results are shown in Figure 5.7. We start by observing the context in which each topic is used, specifically those that apparently relate to similar concepts, or sub-concepts. There are only two contexts in which the generic locations topic is used, when listing locations on the body that hurt (Q1) and when reflecting on the causes of pain (Q5), both with similar percentage of importance, however with great difference regarding the importance of specific locations. This observation tells us that, first, there are indeed references to vague locations on the body that hurt, which may be associated to groups of patients with similar unspecified outlooks on the pain or with specific pathologies that manifest differently in terms of location, and, second, that some people associate cause of pain with source of pain (the wording of question 5 may also have influenced some of the answers). The topics of impacts (1,2) are apparently used interchangeably throughout the whole interview, which makes it hard to reason on their distinction without further exploration.

Finally, we use this figure to try to explain each aspect of the experience of pain by the observed mixture of topics. The location of pain (Q1) suggests being a listing of the locations of the body that hurt. Pain sensation attributes (Q2) are described in many different dimensions, with special attention to location and limiting actions. Pain intensity (Q3) is described in terms of pain activity rather than intensity alone, heavily conditioned by treatment, even though it wasn't suggested in the scripted question. The impacts on the daily life (Q4) are, expectedly, described in terms of the limiting factors, as well as attributes of time. The causes of pain (Q5) are associated with holistic reasoning, reflections, and actual bodily sources of pain. The evolution of pain conditioned by the treatment (Q6) is, expectedly, heavily focused on the treatment. Finally, the expectations of evolution of pain (Q7) are described in terms of pain activity

and treatment. With these observations we conclude that an experience of pain encompasses many different aspects, with each aspect being a mixture of ideas, which further highlights the difficulty of interpretation of our data.

In Figure 5.8a we can observe the distribution of importance of each topic in the population. This observation allows us to understand exactly how important each topic is, and where the majority of the population lies in each topic. If all topics were equally important, or, in other words, none was especially important, each topic would have an importance of around 8%. For our population, we observe three distinct groups of topics, regarding their distribution of importance over the population. The first three topics present a wide distribution, with the largest variance and median values, and only a slight skewness to the right. The following six topics display tighter distributions, with the median value close to that of a uniform distribution, and a heavier skewness. The final three topics, on the other hand, are completely skewed to the right, with median values close to 0% and the third quartile lower than the 8% mark. We also observe that no topic ever has more than 50% importance to a patient in the collection, which adds to the idea that an experience of pain is rarely uni-dimensional.

The aspects tackled by each question are transversal to every experience of pain, and are incentivised to be discussed in the interview. However, the relevance of each of these aspects to each patient, and, thus, encompassing semantical topics, is what shapes their perception of the experience and the description. In Figure 5.8b is plotted, for each topic, the mean importance given by the population, or, in other words, the population's mean mixture of topics, representing what, in general, is more and less important for a patient in our population describing an experience of pain. We define that a topic is relevant for a patient if its weight is more than 8% of the total patient mixture weights. We can observe a clear elbow for this value of importance in both previous figures, distinguishing the topics of treatment, activity, and specific locations from the remaining.

To explain these differences in importance, we raise the following two hypothesis: (1) the design of the interview is such that the three first topics are more incentivised to be discussed than the remaining, artificially suggesting that these are more relevant than the others, and vice-versa, and (2) all aspects of the experience were equally incentivised by the interview design, but some are simply more commonly relevant, and patients discuss them even when not prompted. We refute the first hypothesis by noting that there is exactly one question which

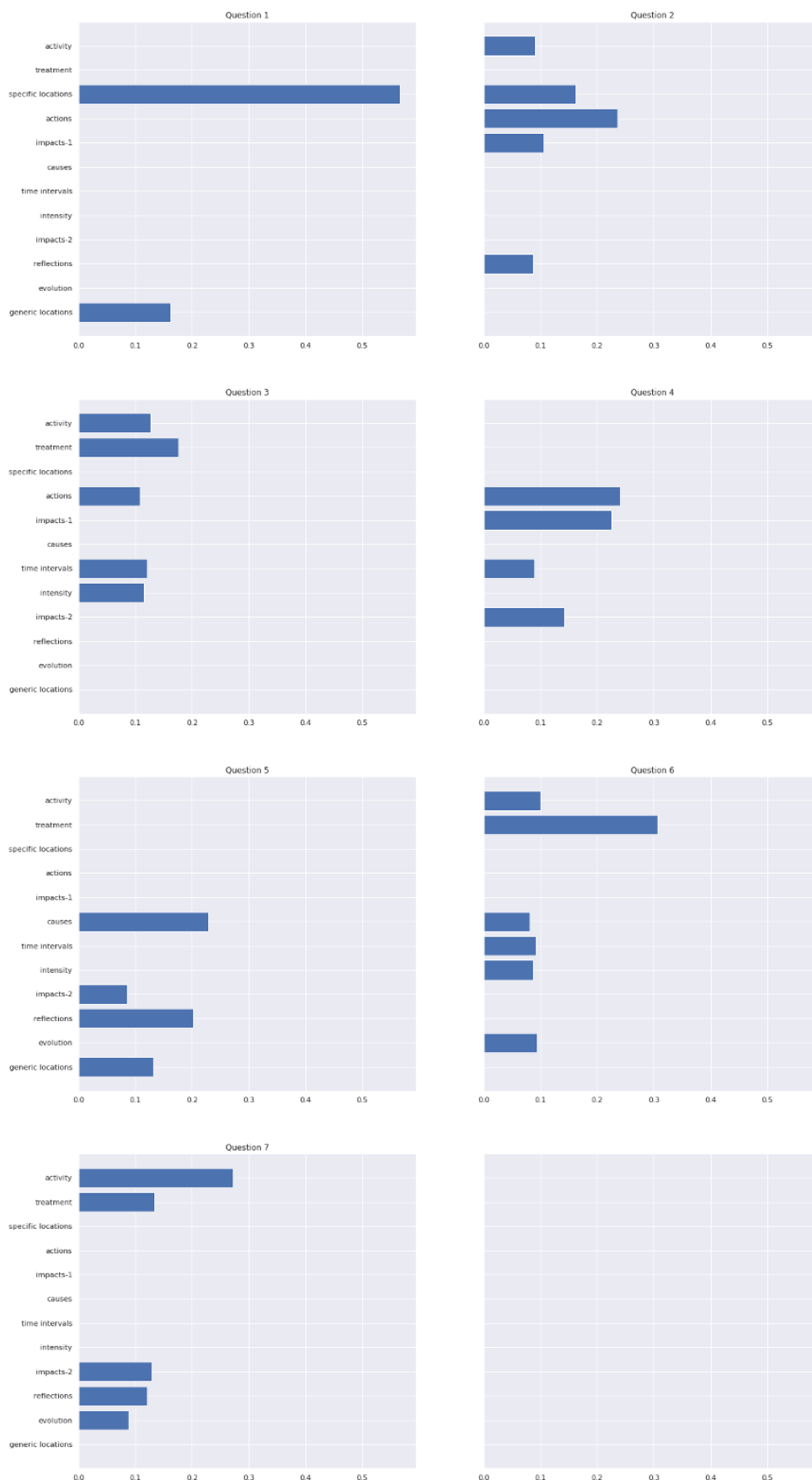


Figure 5.7: Mean topic mixture by question, in percentage. Filtered out importances lower than 8%, for ease of interpretation.

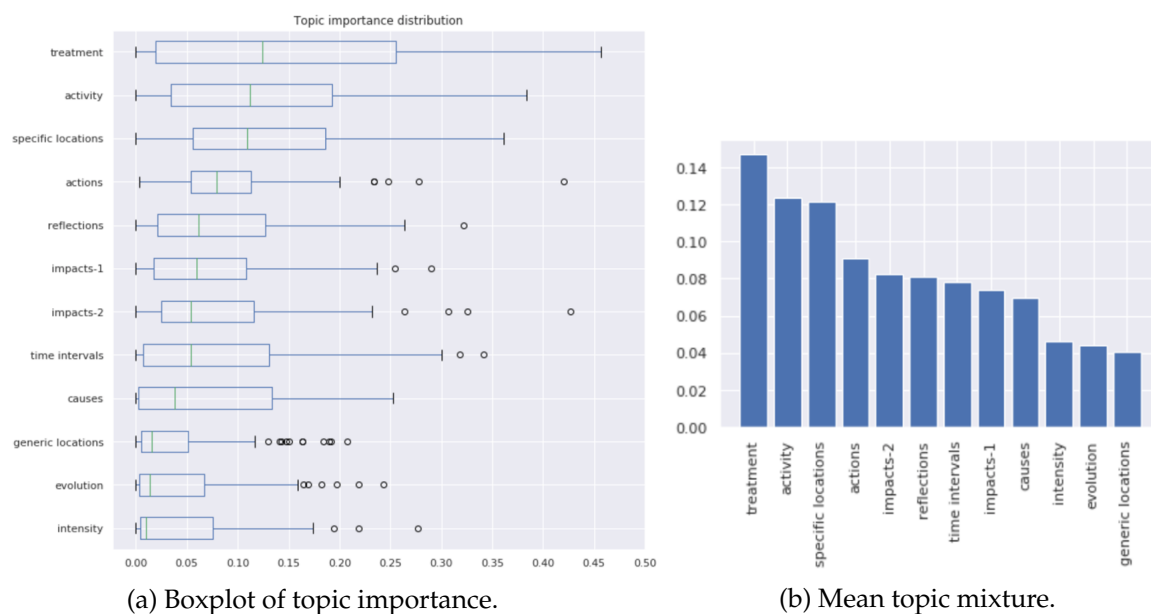


Figure 5.8: Topic importance metrics for the whole population. Importance values are given as a percentage.

prompts the treatment impacts and another that relates to evolution or expectations regarding future developments of pain. However, the treatment topic is relevant for more than half of the population, whilst the evolution topic is marginally relevant. On the other hand, and by referring back to Figure 5.7, we notice that, even though the locations of pain are only prompted once, the corresponding topic is present in the answers of many other questions. The same applies to the treatment topic. This evidence suggests that, for our population, there are aspects of the experience that patients are more commonly inclined to discuss, some without prompting.

In Figure 5.9, we observe the co-occurrence of top 5 most weighted topics of each patient. This allows us to determine which aspects of the experience of pain are commonly discussed together, as if it were a simplistic representation of the train of thought for describing pain. Because our observation is based on the most weighted topics of each patient, and given that a few are overwhelmingly more weighted than others, the observed co-occurrence heat-map with focus on the first 3 topics is expected. According to this evaluation, the most co-occurring topic is activity, especially with the topics of specific locations and treatment. Expectedly, according to our metric, the topic that least co-occurs is the one which most patients (third quartile) find irrelevant, the generic locations topic.

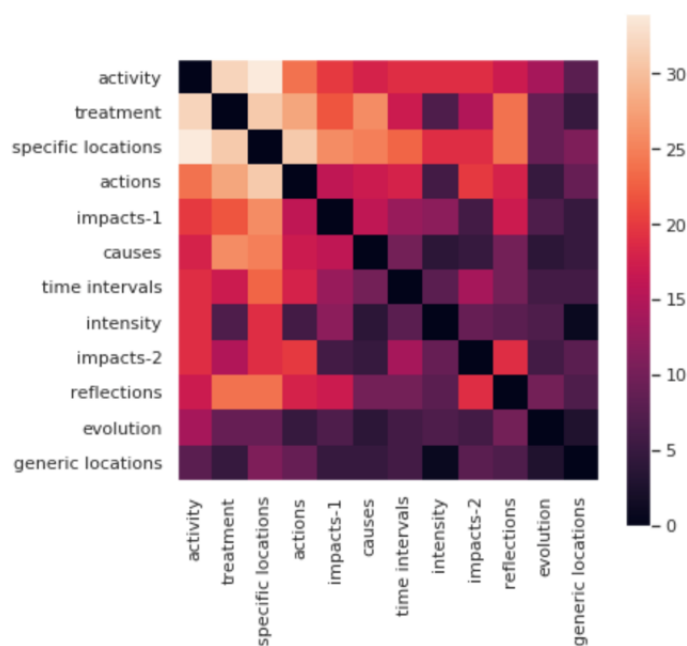


Figure 5.9: Topic co-occurrence in the whole population (top 5).

### 5.3.2.2 Topic similarity clusters

We now define groups of patients that describe their experiences of pain similarly in terms of mixtures of topics. We associate each of these groups with a type of experience, based on the mixture of the cluster, and search for correlations with demographic and clinical parameters.

In the previous section, by analyzing the whole population in the topic space, we determined that the mean mixture of topics mainly focuses on a select few topics, with most topics displaying a very tight distribution. By observing the topic distribution of weights over the population, we assessed that the topics of activity, specific locations, and treatment showed more weight variance than the remaining. This means that if we were to project the patients along those dimensions only, we would be able to better distinguish them into groups than if we considered any other dimension, because, in that case, the patients would be clustered together around the same weight (as observed in the distribution plot). For these reasons, we decide to use these specific topic dimensions to find clusters of patients, in this case with the K-Means clustering model (Hartigan & Wong, 1979).

In Figure 5.10 we can observe the values of inertia, Silhouette Coefficient, Calinski index, and Davies score, across a range of clusters, for this model. We evaluate all these scores because there is no obvious arrangement of patients in well-defined clusters. The inertia value indicates

the sum of squared distances of samples to their closest cluster center. The optimal score is zero. However, as the number of clusters approaches the number of samples, the total sum of distances gets closer to zero, by definition, because in the limit case every sample is its own cluster center. Thus, the ideal number of clusters is obtained not by minimizing the inertia, but by identifying an elbow in the inertia plot, indicating that from that point on, increasing the number of clusters does not result in significant gain in terms of sum of distances. The silhouette score plotted in this figure refers to the mean silhouette of all samples in the dataset, which is calculated as defined before in Eq. (5.2). The optimal silhouette score is obtained when  $a = 0$ , or in other words, when all samples of every cluster fall in the corresponding cluster center. Thus, the closer the silhouette score is to one, the better organized are the samples per cluster. The Calinski score assigns a greater score to dense and well separated clusters, by taking into consideration the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion, for every cluster. In this case, dispersion is measured as the sum of distances squared. Finally, the Davies score is defined as the average similarity between each cluster and its most similar, in such a way that a score closer to zero relates to better separated, dense clusters. All of these metrics favor convex, isotropic clusters, because they are largely based on cluster density, mean distances to centroids and centroid separation. Given these metric plots, we decide to cluster the patients in 7 groups, because it is both a local minimum in the Davies score and a local maximum in the Silhouette, suggesting that this number of clusters finds an adequate arrangement of patients into somewhat dense and well-separated clusters.

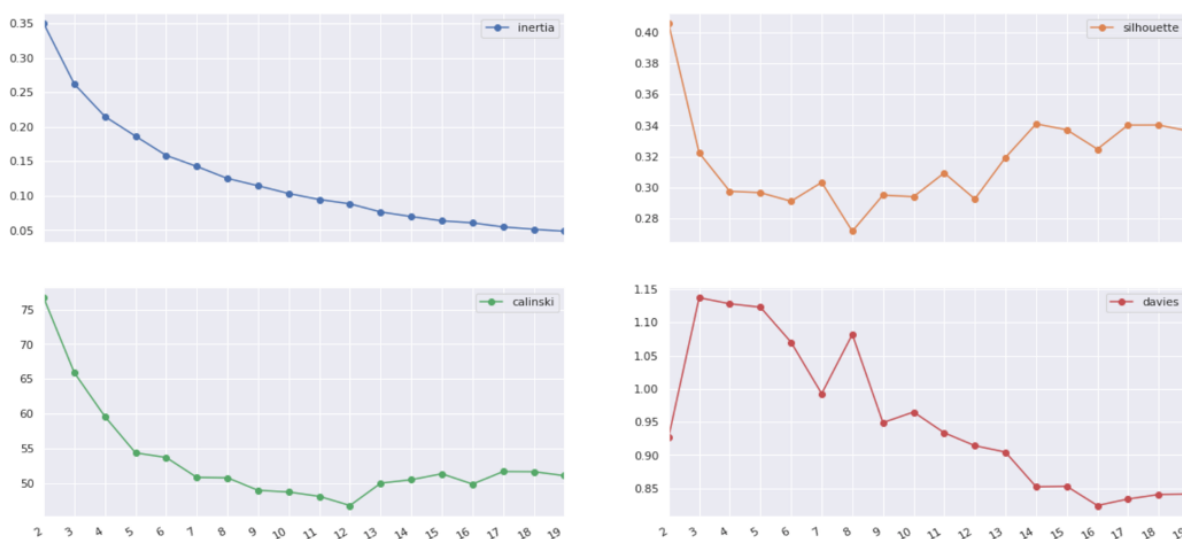


Figure 5.10: Projected documents clustering metrics for a varying number of clusters.



The corresponding projections with PCA and t-SNE (Maaten & Hinton, 2008) can be observed in Figure 5.11. Because the PCA projection tries to explain the distribution of samples in terms of feature variance, maintaining the global structure in sacrifice of sample neighborhood, when the PCA plotted clusters are clearly separated, in general, it can be taken as solid grounds for justifying and explaining said clusters, however, because we are dealing with high-dimensionality data, the groups may be clearly separated in a high-dimensional space and be plotted one over the other in PCA-dimensions, possibly resulting in overlap as can be observed. The t-SNE projection is plotted in parallel, to assess the overlaps observed in the PCA projection. The t-SNE projection tries to preserve sample neighborhood in high-dimensionality in a lower-dimensionality, and, thus, can obtain a more informative view of the clusters behavior in high-dimensionality. With these plots, we observe that some clusters are clearly separated from the rest, whilst others have considerable dispersion. This is an expected observation from the metrics discussed before. Overall, we conclude that this is an adequate arrangement of patients in clusters, according to their projections in the topic space.

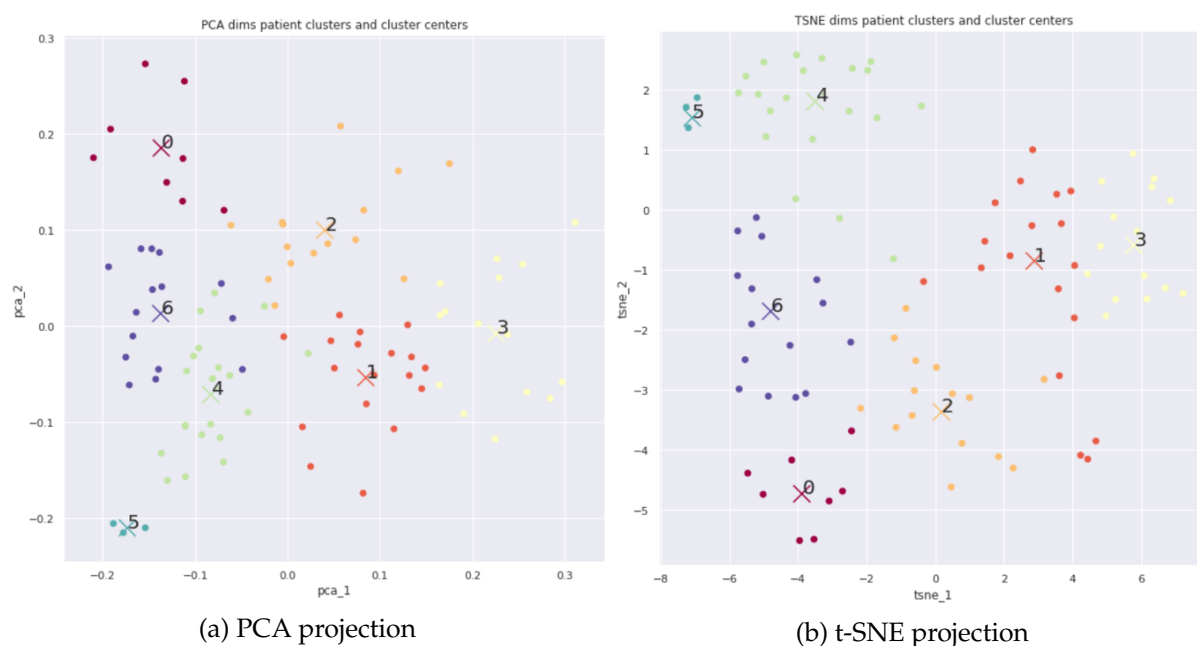


Figure 5.11: Projected documents on a 2D visualization, with color codes referring to the assigned cluster, for a total of 7 clusters.

Observing the mean topic mixture of each cluster, we can gain insights as to what each might represent. This is shown in Figure 5.12. Because we are assuming the mean topic mixture is representative of all samples in a cluster, the following discussions disregard Cluster 5,

since the corresponding number of samples is very reduced (3 samples), and there is no obvious separation of these clusters from the remaining in the projection plots in Figure 5.11. Indeed, these samples could very well be outliers of other clusters. We observe that the mean mixtures of our clusters are characterized by a high weight given to one or two topics and small weights scattered across other select few topics (weights below 8% are filtered out). As expected from previous results, the topics that are most commonly assigned high weights, and are used to somewhat easily distinguish each cluster (with the exception of Cluster 4), are the ones that presented more variance and patient counts, the topics of activity, treatment, and specific locations. Figure 5.13 shows the distribution of all topics in each cluster. Specifically focusing on these three topics, we can observe that each cluster assigns a specific range of importance to each, allowing for a clear distinction between these groups of patients in this section of the topic space. This is an expected result, since the clustering was done considering only the dimensions of these topics.

At this point we have obtained groups of patients that have similar descriptions of their experiences of pain, as given by their mixtures of topics. Figure 5.14 shows the distribution of the continuous parameters, specifically, age (demographic), duration of the disease, pain intensity, ESR, and RCP (clinical), per cluster. We do not observe any significant difference of distribution of any parameter in different clusters (again, disregarding Cluster 5 due to its small size). In Figure 5.15, we can observe the distribution of the categorical parameters per cluster, specifically, pathology (clinical), whether it's an active person professionally, level of education, and sex (demographic). Noticeably, most patients are diagnosed with either E or AR, the majority has a primary level of education, and most are of the feminine gender. All other parameter categories are residual and, thus, not comparable. Again, these observations suggest that there is no correlation between the obtained types of experiences of pain, according to their mixtures of topics, and these parameters.

### 5.3.2.3 Demographic and clinical clusters

In this section we group patients by values or ranges of demographic and clinical parameters, as given by Tables A.1 and A.2, and Figure A.2. For each group of patients, we consider the mean topic mixture to be representative of the patients in that group. We consider a group of patients to have a specific experience of pain if its mean mixture of topics can be well differen-

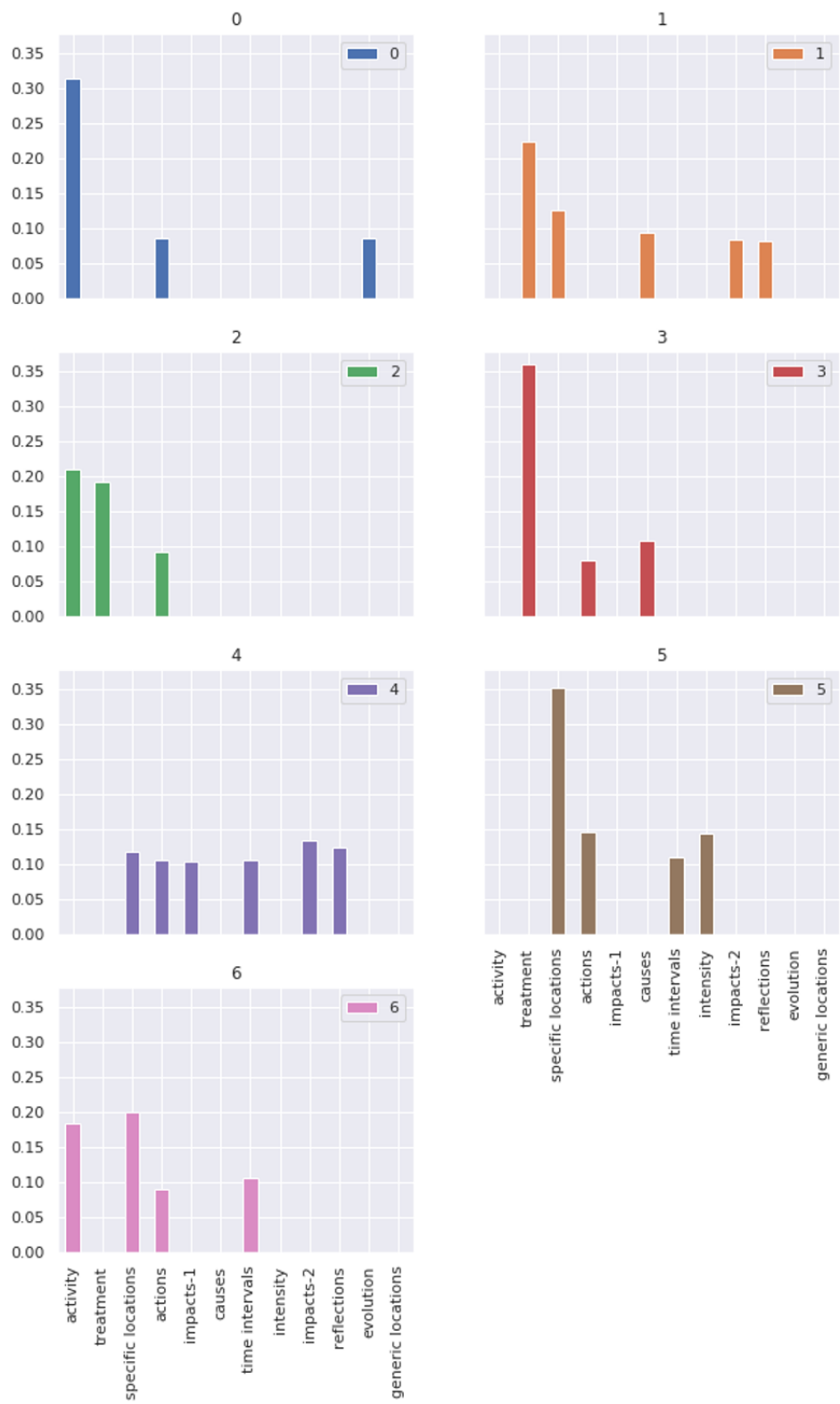


Figure 5.12: Mean topic mixture of each cluster of patients.

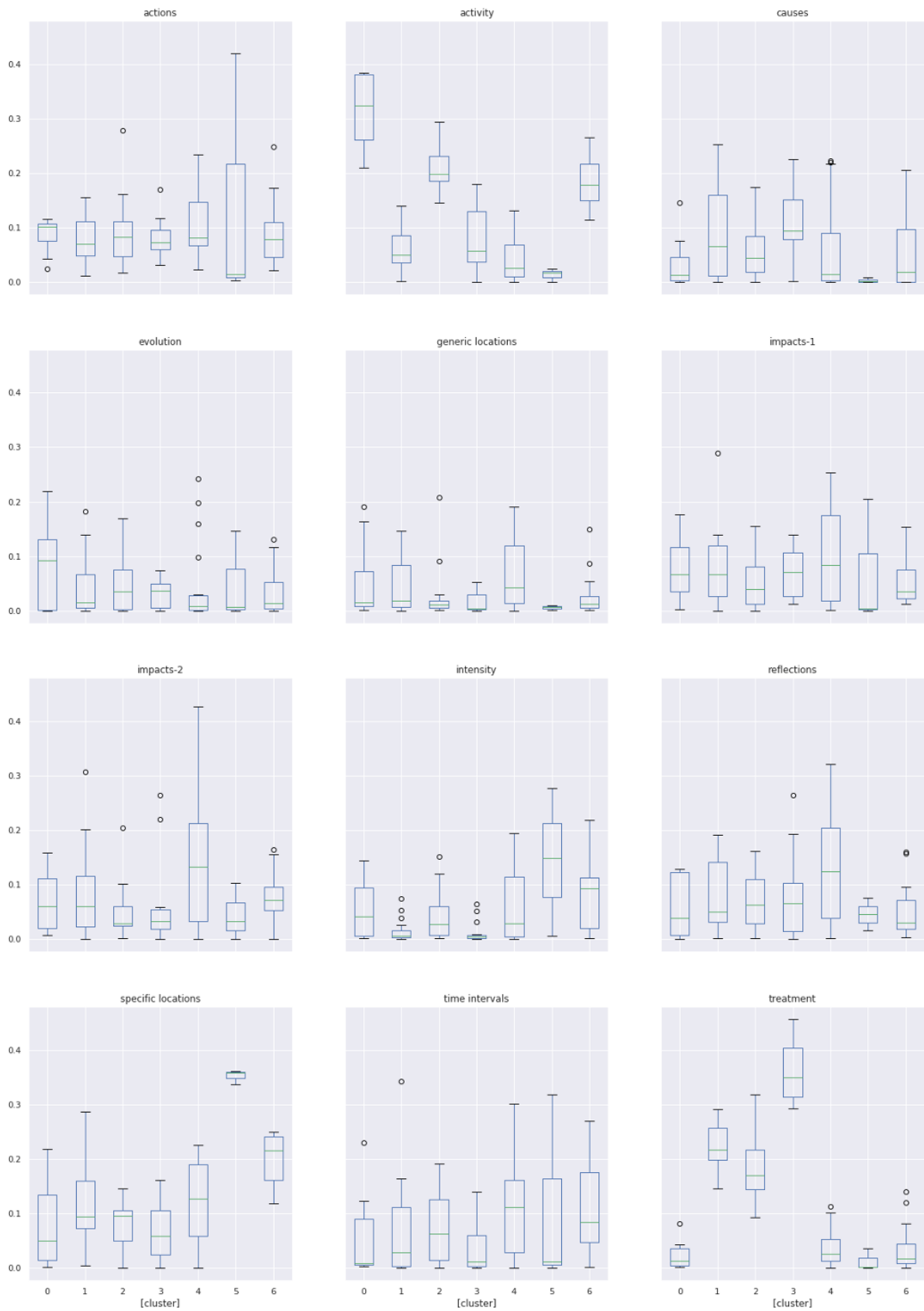


Figure 5.13: Topic importance distribution by cluster.

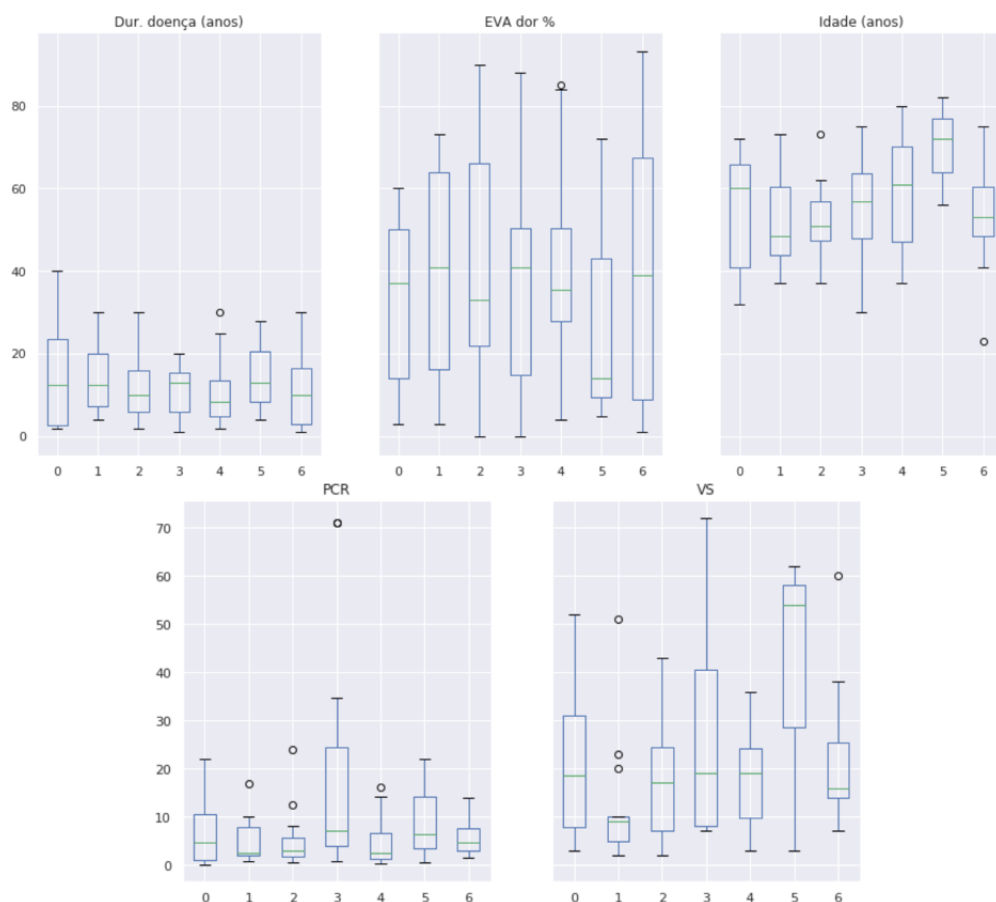


Figure 5.14: Distribution of continuous demographic and clinical parameters on each cluster of patients.

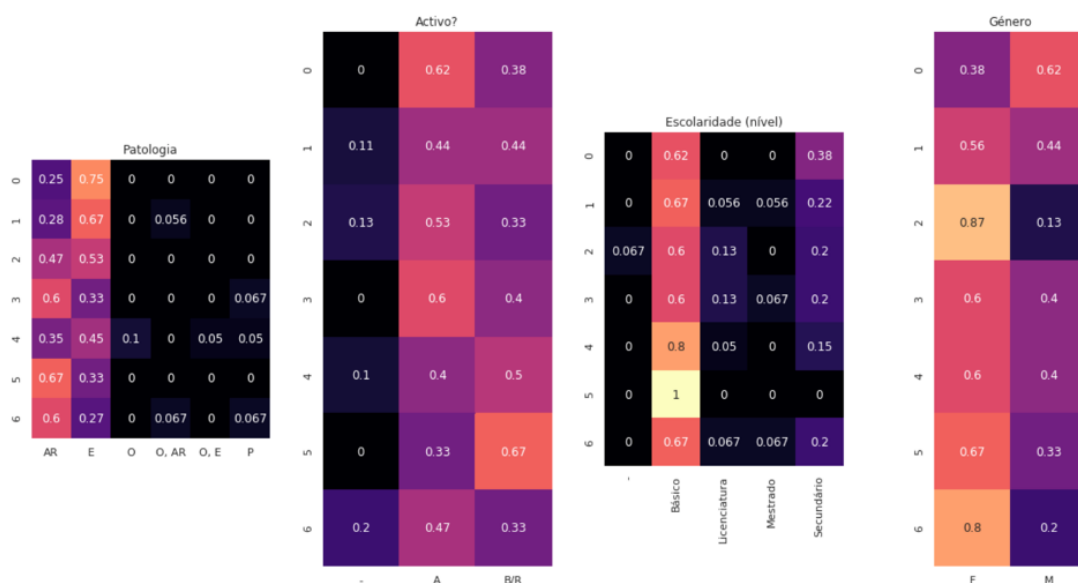


Figure 5.15: Distribution of categorical parameters per cluster (rows sum to 1).

tiated from the mean mixture of the whole population. The results for each group are shown in Figure 5.16, as well as the mean mixture of the whole population, repeated here for ease of discussion. The results of the parameters that did not show relevant results for discussion, specifically, sex, duration of the disease, ESR, and RCP, were omitted. We can observe in this figure that all groups of patients follow, in general, the mean mixture of topics of the whole population, with a few notable exceptions.

Starting with clinical parameters, the group of patients diagnosed with Spondylitis (E) differs from the group diagnosed with Rheumatoid Arthritis (AR) mainly on the topics regarding the locations of pain. This observation is not unexpected, since different pathologies may have different manifestations of pain, including different and more or less specific locations. In this case, we observe that for pathology E specific locations of pain are less important (as given by the weight of the topic), than generic locations. The contrary observation can be made for pathology AR. Between these pathologies, the other relevant differences lie on the topics of causes and impacts (2). One possible interpretation is that because AR is associated with more specific locations of pain, the corresponding patients can more easily associate a cause to the pain than the remaining. A similar interpretation may be applied to the patients diagnosed with E. Because this pathology is associated with more generic locations of pain, it may have more impacts on the daily life of the corresponding patients than the remaining. Observing now the groups of patients as given by the levels of self-reported intensity of pain, we make the following remarks. The very similar mean topic mixtures of the groups with pain intensity [0-25] and (25-50] suggest that these patients have similar experiences of pain, which does not apply to the remaining levels of intensity. The group which reports the highest level of intensity is clearly distinct from the others, showing a lot of emphasis on the specific locations, actions, and time intervals of pain activity. However, this distinction can be associated with the unbalancing of the groups.

Now focusing on the demographic parameters, we do not observe as notable differences as with the clinical parameters. Indeed, the patients with ages comprised in (20-40] give less importance to the causes, intensity, and specific locations of pain, although that may be correlated with the group being mainly diagnosed with E, due to the higher importance given to the generic locations, in comparison to the other groups. The patients with the lowest level of education (Básico) show a much higher importance in the specific locations of pain than any

other group, and the patients with the highest level of education display a similar behavior but for the causes topic. This difference, however, may be associated with the unbalancing of the groups. Finally, we observe that the patients who are not professionally active (retired or on medical leave) are more focused on the impacts of pain than the ones who are active.

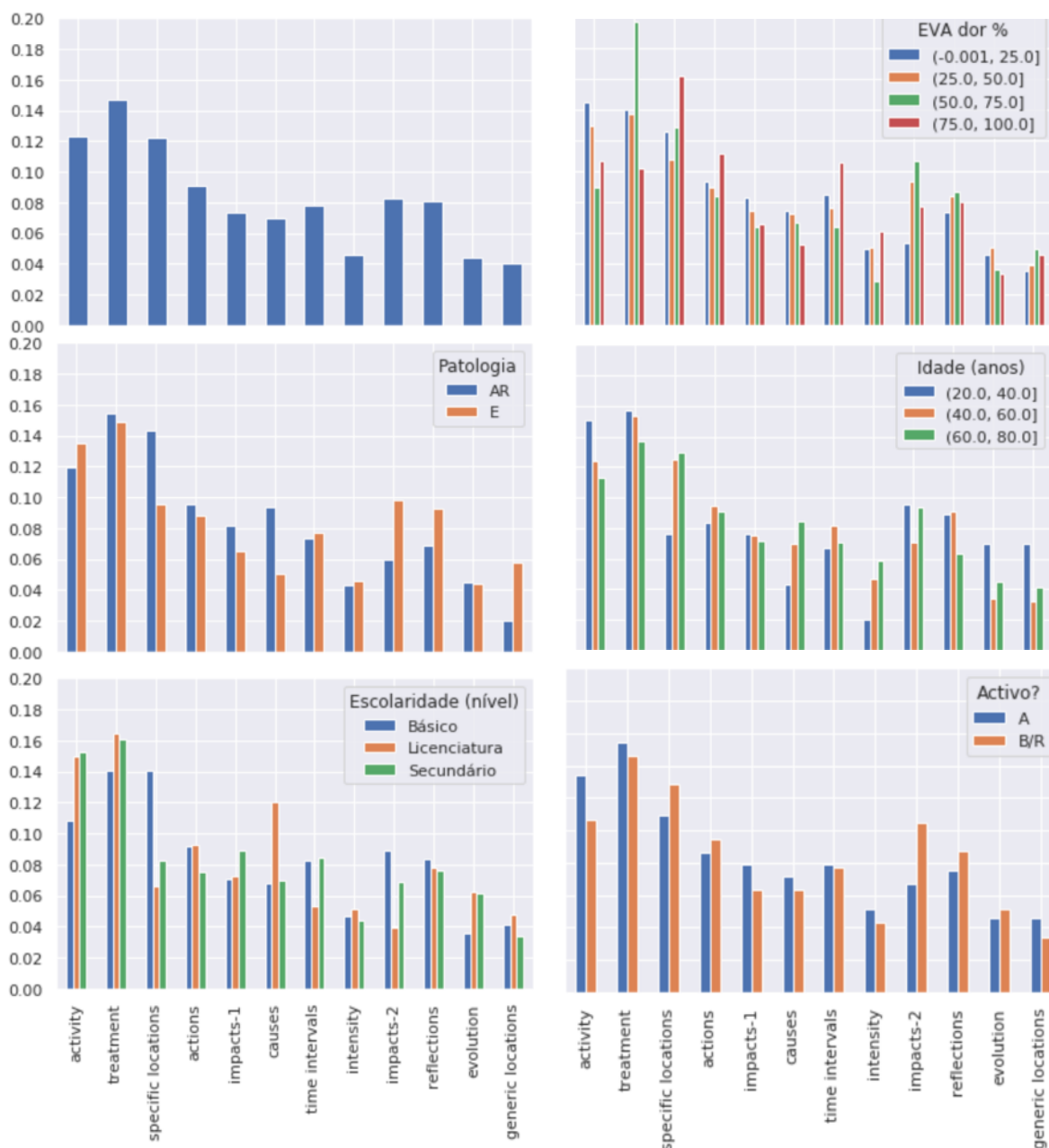


Figure 5.16: Mean topic mixture by group of patients, per clinical and demographic parameters.

## 5.4 Summary

In this chapter we propose to project and characterize the population in a linguistic feature space, in order to relate semantical aspects of descriptions of pain with intrinsic qualities and extrinsic parameters. This is done in terms of topic mixtures.

We started by evaluating different topic models on the fragmented documents, so that the most adequate topic space could be determined. In this discussion, we have observed that both the limited amount of documents and their short-text characteristics do not allow for coherent topic extraction of statistical inference-based models and simple word co-occurrence based models, as evidenced by the behaviors of both baseline models, LDA, and NMF. We have also observed that even though the extracted topics are highly coherent according to the PPMI metric, the resulting SeaNMF topic space has a poor distribution of documents and displays high overfitting. We have associated this behavior with the limited amount of documents. Finally, given the interpretability and clustering metrics, we conclude that topic modeling with CluWords (FastText), after applying the preprocessing pipeline, produces the most adequate overall representation of the collection of documents in the topic space, which is in agreement with the increased value of using external semantic information to deal with short-text documents, and softening the sparsity of the vocabulary space, even though this is limited by the richness of the vocabulary (which has been shown to be very limited).

Following this evaluation, we used the CluWords (FastText) topic space to characterize the population, under three approaches. The overall population analysis laid the foundations for further, finer analyses, with the identification of the topics which were more relevant for the population, and that best differentiated the population. Then, groups of similar descriptions of pain were identified, as defined by their mixtures of topics. We inferred the semantics associated with each of these groups as representing distinct types of experiences of pain, and found no correlation with demographic and clinical parameters. Finally, we studied the hypothesis that the different demographic and clinical parameters influence the perception, and consequent description of the experience of pain. To this end, we split the population into groups as given by these objective parameters, and evaluated the difference between these groups in the topic space, according to their mean topic mixture. Relevant differences were identified and discussed, although the majority displayed similar distributions.



# Prediction of Clinical Parameters

In this chapter we raise the hypothesis that expressions of pain, specifically, verbal descriptions of chronic pain experiences, convey potentially useful information to aid in the assessment of clinical parameters of rheumatologic patients. This suggests that there is a direct relation between the linguistic manifestation of pain (a description of the experience) and the clinical parameters of the corresponding patient. The methodology employed to study this hypothesis is that of a prediction task, with features extracted directly from documents of pain descriptions. The following sections describe the setup and evaluation associated with the extraction and use of these linguistic features from the intrinsic semantical structures of the descriptions of pain, based on the previous chapter that characterized the population in that domain.

## 6.1 Task definition

The task proposed in this chapter may be performed on any parameter. In this case, we are interested in both the diagnosed pathology and the reported intensity of pain. Both of these parameters are directly related with the experience of pain, even though the design of the interview, which is the method used to collect descriptions of pain from patients, was not directly designed for this task. The same set of features and evaluation is used for both parameters.

### 6.1.1 Pathology classification

Patients are distributed per pathology as presented in Table 6.1. Given the poor distribution across all classes, this experimental setup is only concerned with P1 (Rheumatoid Arthritis) and P2 (Spondylitis), so that the task is defined as binary classification.

P1	P2	P3	P4	P5	P6	Total
41	45	2	2	1	3	94

Table 6.1: Distribution of patients per pathology

### 6.1.2 Pain intensity classification

Patients reported pain intensity through a VAS, as presented in the data collection form in Figure 4.1. The used visual scale is of 100 mm, where the left endpoint represents 0% pain, and the right endpoint represents 100% pain, which is the maximum pain ever experienced by that patient. The collected data allows for a fine regression task, on a 0-100 scale of pain intensity, or a more coarse task, by down-sampling the values to a 0-10 scale. However, at the time of writing, the number of instances limits these possibilities. Therefore, we define a multi-class classification task, by separating intensity values into four levels, as defined in Table 6.2. Given as there are more patients that report less pain, the classes are very unbalanced, and this is expected to skew the results.

[0-25]	(25-50]	(50-75]	(75-100]	Total
29	31	18	8	94

Table 6.2: Distribution of patients per level of pain intensity.

## 6.2 Feature extraction

To each patient is associated a collection of 7 documents corresponding to the transcription of each question's answer. The linguistic features for each document are extracted as described in Chapter 5, and summarized in Table 6.3. The first 4 features are the baselines. The vocabulary-based representations (BoW and TF-IDF) are introduced so that the gain in using topic modeling may be assessed. According to these features, each patient is associated with a group of 7 vectors, either of dimension  $V$  or  $k$ . In order to represent each patient with a single vector, the following types of aggregation are considered, *fragment*, *patient*, and *single question* [1-7]. These are explained in Table 6.4.

The *fragment* aggregation looks independently at each of the 7 documents belonging to a patient, as if they were not semantically related. This way, both vocabulary and topic space representations of the collection are completely different from the other aggregations, given that there are more independent documents, and each is much shorter in length and semantically focused on less topics.

The *patient* aggregation considers that each patient has a single, long, document (the result of concatenating beforehand all 7 fragments for each patient). This means that both vocab-

Features	Dimensions
BoW	$D \times V$
TF-IDF	$D \times V$
LDA	$D \times k$
NMF	$D \times k$
SeaNMF	$D \times k$
CluWords (FastText)	$D \times k$
CluWords (BERT)	$D \times k$
BERT (doc2vec)	$D \times k$

Table 6.3: Considered types of features to extract from a document collection.  $D$  is the number of documents in the collection,  $V$  is the size of the vocabulary, and  $k$  is the number of extracted topics.

Identifier	Aggregation method description
<i>fragment</i>	The 7 vectors are aggregated by their mean value in each dimension.
<i>patient</i>	The 7 documents are concatenated beforehand, resulting in a single document per patient which is then transformed into a vector by the methods above.
<i>single question [1-7]</i>	Only one of the vectors is considered (corresponding to a single question of the interview).

Table 6.4: Possible aggregations that represent each patient by a single vector.

ulary and topic extractions are now applied on only 94 documents (equal to the number of patients), albeit richer and longer. However, given that the number of documents is so low (compared against the original 656), there might a loss of information, especially regarding word co-occurrence in documents and complex topic distributions. These problems could be solved by increasing the number of patients, which cannot be done artificially. For these reasons, the results associated with this type of aggregation are expected to be inferior to that of the *fragment* aggregation.

Finally, the *single question [1-7]* aggregation presupposes that for the current task, the patient is sufficiently, and better, represented by a single question’s answer to the entire interview, since there is much less noise and the text is semantically focused. In this case, the number of documents is also reduced to the number of patients, however taking a big cut off the collection’s vocabulary. If, in fact, there are question’s answers in the interview which are prejudicial to the prediction of the associated clinical parameter, or are simply irrelevant, diluting the useful information in noise, this type of aggregation is expected to produce superior results.

Finally, in order to understand the relevance of each question in the interview for the clinical parameter classification task, all experiments are done in an ablative fashion. This way, each experiment includes all possible permutations of the considered interview questions. The motivation for this ablative approach is the intuition that certain questions in the interview actually provide enough information for the parameter classification, and that the use of the answers to certain questions can introduce noise, which may negatively affect the result.

### 6.3 Evaluation

Given the limited size of the dataset, it is not separated in training and test sets. Rather, the evaluation is performed following the Leave-One-Out method, so that the result of each experiment is the mean accuracy score of training on every subset of  $n - 1$  patients and predicting the clinical parameter of the remaining. All experiments are evaluated according to their accuracy.

### 6.4 Results and discussion

Parameter	Values
Text type	[natural, lemma]
Stop-words	[remove, not remove]
$\alpha$ -CluWords (FastText)	0.55
$\alpha$ -CluWords (BERT)	0.98
$k$ (number of topics)	12

Table 6.5: Text parameters of the experiments.

	Type of text	Stop-words
Exp. 1	natural	not remove
Exp. 2	natural	remove
Exp. 3	lemma	not remove
Exp. 4	lemma	remove

Table 6.6: Configuration of all experiments.

In this section we present and discuss the results associated with the previously presented tasks. The type of text used for feature extraction and further analysis can have a great impact on the results. Thus, the text parameters that we are interested in studying, specifically to understand their influence on the quality of the prediction, are summarized in Table 6.5. Given

these, there is a total of 4 experiments, which are all permutations of the parameters in this table, summarized in Table 6.6. Each experiment encompasses the accuracy score of 8 feature types, across the 9 types of feature aggregation.

The experimental setup relied on the use of 4 machine learning models. These are Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Random Forest (RF), and Logistic Regression (LR). The hyper-parameters of each model are used as defined by default in the Sci-Kit Learn toolkit, except for the random seed, which is fixed for reproducibility. Instead of using only one, all of these models were chosen due to this being an exploratory study, therefore interested in minimizing the variability of results, especially those that are intrinsic to the design of the models. After running the experiments for all of these models, it was determined that the performance of all models was equal, or inferior, to that of the SVM. For this reason, all results and considerations shown here are in regard to the SVM model with a linear kernel.

#### 6.4.1 Pathology classification

Figure 6.1 shows the mean accuracy score per experiment configuration. It also represents the score variance per experiment, and a red dashed line representing the threshold of random choice (50%). This plot allows us to compare experiments in a high-level and to understand the limitations of each aggregation type, in general. Observing this figure, we first conclude that the variance in score for each experiment, across all aggregations, is very small, which validates this type of aggregate observation and the following discussion about the general performance of each experiment, irrelevant of the feature types. Indeed, there is a clear trend in monotonicity across all experiments, which allows us to attribute the justification to the aggregation types, rather than the experiment configuration. The *fragment* type displays higher scores than the *patient* aggregation type, even though not as relevant as expected. By aggregating the 7 vectors by their mean value in each dimension, we are considering all documents to have the same importance to the general representation of the patient, which is not necessarily true, and might be the cause for information loss. A possible approach to overcoming this is to weight each vector, given the importance of each question to the task of pathology classification. Determining these weights is not a trivial task. We can also observe a clear spike in accuracy, for all experiments, when using the *single question (1)* aggregation type. This means that the patient answer's to the question "Where on your body does it hurt?", is informative enough to

predict their pathology in our binary classification setting, with a mean accuracy score above 70%. This result is in line with the clinical literature, where Rheumatoid Arthritis is said to typically manifest with pain on multiple, scattered, joints, more commonly on the extremities, such as the wrists (Rindfleisch, Adam J and Muller, Daniel, 2005), opposed to (Ankylosing) Spondylitis, which is commonly associated with inflammatory back pain (McVeigh & Cairns, 2006). Basing the prediction only on answers to questions (2), (3), (4), (6), or (7), yields results similar, or inferior, to random binary choice. Finally, the answers to question (5) also seem to allow for prediction results comparable to the *fragment* and *patient* aggregation types.

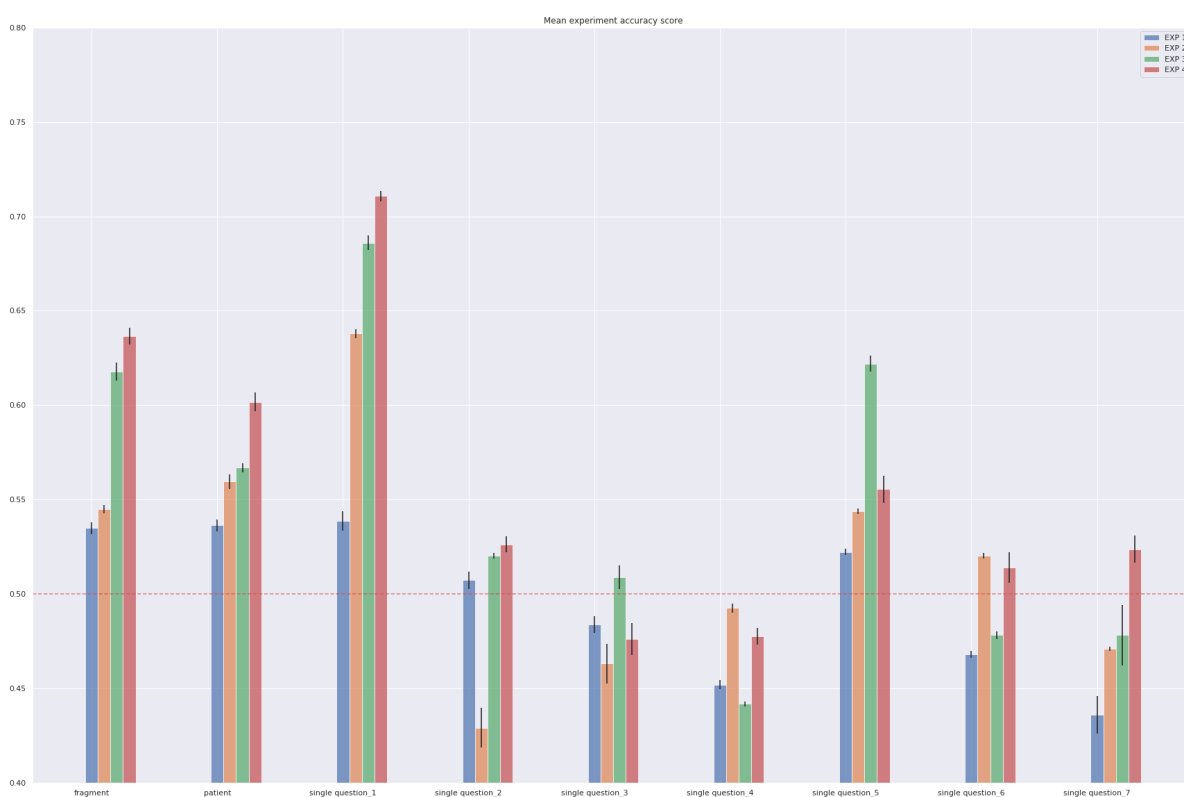


Figure 6.1: Mean accuracy score of each experiment in Table 6.6, over the different types of feature aggregation in Table 6.4.

Focusing on the relevant aggregation types (*fragment*, *patient*, *single question (1)*, *single question (5)*), we can now compare the performance between experiment configurations. We conclude that EXP. 1 results in the poorest performance overall, which can be justified by the fact that it is based on the most raw data (original next, no stop words removed), meaning that important information gets diluted in noise. This is especially evident for *single question (1)*, which is basically a list of nouns (locations on the body), where the mere presence of syntactic building blocks of words, such as determinants, pronouns, and conjunctions, and the syntac-

tic variability of words, may dilute the information carried by the relevant nouns, resulting in a performance score more than 5 percentage points inferior to the remaining experiments. Finally, even though there is some evidence that, overall, using lemmatized text (EXP. 3, 4) results in better accuracy scores, the gain is not as large as expected. The removal of stop-words is also reflected in the small difference between EXP. 3 and 4. The following discussion will focus only on these two experiments.

Figures 6.2 and 6.3 dive into the actual scores, per experiment, per feature type. These plots allow us discuss which types of features seem to be more adequate for the defined task. In Figure 6.2, looking at the vocabulary-based baselines of feature types (BoW and TF-IDF), we observe a significant difference when not removing (Figure 6.2a) and removing stop words (Figure 6.2b). Specifically, by removing stop words, simple word frequency and co-occurrence given by the BoW features is as informative as the TF-IDF features. As expected, given that the text is already standardized (lemmatization), the TF-IDF features are capable of extracting important information, regardless of having or not removed stop words, because these are usually assigned very low scores due to their high document frequency nature. Shifting our focus to the baseline topic-based features (NMF and LDA), on the same experimental settings as before, we observe a clear distinction in favor of the NMF model. In fact, LDA accuracy scores are as good as, or worse, than random choice, in most cases. Recalling the discussion about the LDA topic space in Chapter 5, due to the limited amount of documents and the short-text nature of these documents, this is an expected observation. This is also in line with the observations made about the BoW and TF-IDF features (LDA is limited by the information carried by the BoW representation, and NMF is limited by the information carried by the TF-IDF representation). Both BoW and TF-IDF features present higher accuracy scores than NMF, overall (in some cases, with an increase of almost 20 percentage points). However, it is important to note that when referring to the *fragment* aggregation type, there is no evident distinction between these 3 types of features. Indeed, this suggests that for the task of binary pathology classification, a listing of pain locations is more informative than any other type of observation on the patient's pain manifestation. The same reasoning applies to all other topic models, which scores are plotted against the best baselines in Figure 6.3. Their performance on this task is not evidently different from the baseline NMF. Finally, the doc2vec features, given by a pre-trained BERT word-embedding model, do not seem to produce interesting results. This may be attributed to the lack of adaptability of the pre-trained model to the context of our data.

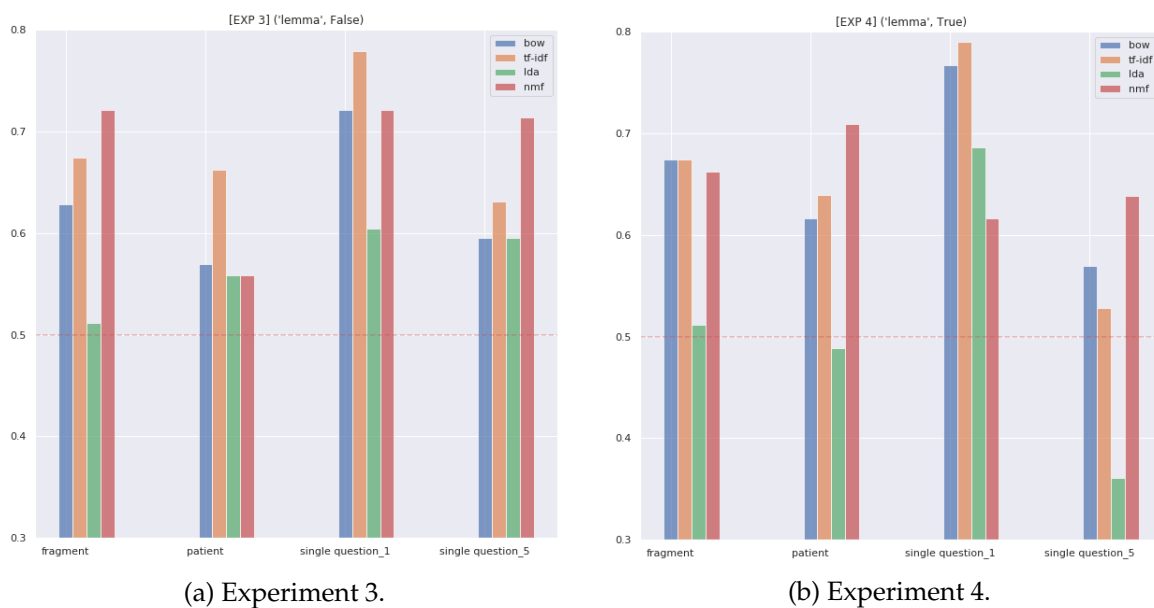


Figure 6.2: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Focused only on the baselines.

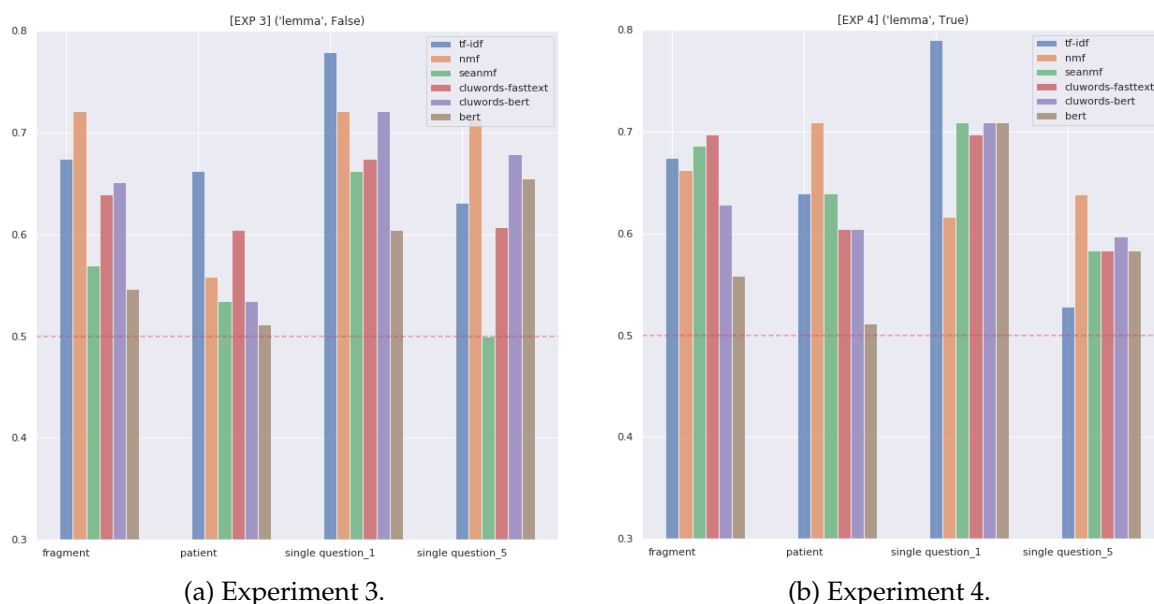


Figure 6.3: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Some baseline results were omitted for ease of read.



With this discussion, we conclude that for our setting of binary pathology classification (specifically, between Rheumatoid Arthritis and Spondylitis), the TF-IDF features are, overall, the best information extraction method, with an absolute score of 79% with lemmatized text and removed stop words (EXP. 4), considering the *single question (1)* aggregation type. This observation is the main motivation behind the ablative experiment.

Finally, Figure 6.4 reveals the mean accuracy score, for all experiments, in an ablative fashion, regarding every permutation of excluded questions from the dataset. This extensive ablative evaluation provides us with insights into how answers to each question in the interview impact the final classification task. We observe a recurring pattern: whenever answers to question (1) of the interview are ignored, whatever other answers are also discarded, the accuracy score decreases significantly, with very low variance across experiment configurations (sometimes going below the random choice level, plotted by the red dashed line). We also observe a slight increase in score as we remove more answers that are not from question (1) (higher scores along the horizontal axis, as suggested by the orange dashed line). This is in line with the previous discussion, and can be summarized by the importance of pain location for diagnosis of these specific pathologies.

#### 6.4.2 Pain intensity classification

The obtained results in this task were equal, or inferior, to random choice (25% accuracy). For this reason the results were omitted, limiting this section to the discussion on why there was no success (results may be consulted in Appendix B). We have identified many possible reasons to explain what was obtained. These are mainly related with data availability and the nature of the data and task.

This task was performed on highly unbalanced classes, and, as expected, after observing the confusion matrix, we concluded that the less favored classes (the last two levels of intensity) were never being predicted. The unbalancing problem by itself should not have such an effect, however, because the number of samples is so low, it can be enough to distort the model's reasoning. On top of this, as we have identified when characterizing the groups of patients according to these parameters in the topic space, the first two levels of intensity, which are also the ones with most samples, are practically indistinguishable. This suggests that the class definition may also be faulty by nature.



The self-reported pain intensity was obtained by means of a visual scale, where the endpoints represent, respectively from left to right, zero pain, and the most pain ever felt by that patient. This means that if a patient were to report 100% of pain, there is no guarantee that that is equivalent to the next patient's maximum pain ever experienced. Even though this is the reasoning behind the grouping of intensities in 4 broad levels, to account for subjectivity error intervals, these results may suggest that indeed the reported intensities are very disparate, subjectively speaking.

Finally, another possible reason for the observed poor results is related with the design of the data collection protocol. Pain intensity reports are instantaneous, meaning that they are mainly relevant in the instant of time in which they were reported. However, the interview was designed to accommodate for past events and evolution of pain, possibly skewing the patients to refer aspects of the experience that are not relevant in that precise moment that they also reported pain intensity. Indeed, there is evidence of patients stating that their description is according to a previous moment in time, rather than the present. Therefore, there may be an offset between the aspects discussed in the interview and the actual report of intensity.

## 6.5 Summary

In this chapter we have evaluated the hypothesis that the manifestation of certain clinical parameters would be reflected on the descriptions of the corresponding experiences of pain.

Specifically, for pathology classification, we analyzed and discussed which linguistic features extract the most relevant information. Additionally, we determined which parts and conjugations of these descriptions actually conveyed enough information for pathology classification above random choice, highlighting the TF-IDF vocabulary-based features based on the answers to the question "Where on your body does it hurt?", with 79% accuracy in the binary classification task.

Finally, we discussed the sub-optimal results associated with the classification of the level of intensity of pain. We identified the different reasons that might be skewing the results, which range from the data availability to the problem definition itself.



# 7 Conclusions and Future Work

The present work explored the computational analysis of the language of pain descriptions, specifically in a healthcare setting. The overview of the nature of pain in Chapter 2 allowed for the characterization of the different experiences of pain and possible causal agents, specifically focusing on the chronic pain experience. By exploring the cognitive process which undergoes this experience, the main cognitive aspects that affect in some way the perception and consequent expression of pain were identified, namely, the emotional state, beliefs, expectations, behavior, and the sociocultural context of the subject. Based on these observations, in Chapter 3, the methodology applied to the linguistic and paralinguistic analysis of similar problems was explored. The method that was identified as the most adequate for the linguistic analysis is topic modeling, tackling the various aspects of the experience of pain previously studied. On the other hand, the paralinguistic analysis was identified to be based on speech modeling, specifically the extraction of acoustic features, to further characterize the descriptions.

## 7.1 *Conclusions*

The following sections present final considerations on Chapters 4, 5, and 6, which dived into the collection and preparation of the dataset, and the presentation and discussion of the experimental setups comprising the analyses.

### 7.1.1 **Data collection**

The data were collected and prepared specifically for the present work. Indeed, there was the opportunity of tailoring the collection for the intended analyses, resulting in the design of the interview and complementary form presented in Chapter 4. Even though the interview did guide the patients to discuss the aspects of the experience deemed most relevant for evaluation, its strict format may have forced some patients to discuss aspects that were not relevant to

them, or discuss them in an way that was not natural. This resulted in some answers being very imprecise, and, in rare cases, with apparent discomfort on the part of the patient. Another consequence of the tailored interview is the fact that it cannot be used to collect a parallel dataset of a control group. Indeed, the very definition of a control group is very difficult.

A possible correction to our approach includes a re-wording of the questions to more grounded terms, so that all patients are capable of understanding the aspects being discussed. Another solution would consist of a change in the approach, designing a single, open question, that would ask the patient to describe the experience however is found fit. This would also encompass the possibility of having a control group, because it could be applied to the description of any other experience. Naturally, this approach would have its own downfalls, including the possibility of having no patient discuss any of the relevant aspects, or in a very vague manner, possibly rendering it void.

Regarding the limitations of the paralinguistic analysis, this would require a more intricate setup for the data collection. The proposed and implemented setup, with the data being collected with a recording smartphone, was intended to, first, not overwhelm the patient, causing further discomfort, and, second, not pressure the healthcare system by overloading the interview with a complicated setup time. A possible solution consists of discarding the importance of the collection being in a healthcare environment, having a proper setup in a location agreed with the patients. However, this approach is expected to greatly limit the number of patients willing to participate, given the possible limitations imposed by the disease.

Overall, even though the obtained dataset has its limitations and challenges, it was possible to perform the intended linguistic analysis with relevant results.

### **7.1.2 Linguistic characterization**

The linguistic characterization of the population, presented and discussed in Chapter 5, consisted of the topic modeling of the collection of documents, and the identification of similar groups and correlation with objective, external parameters.

It was decided to approach the evaluation of the different models in a fragmented way, considering each answer, to each question of the interview, to be independent in terms of latent semantical topics, even though belonging to the same patient in groups of 7 fragments. The

decision was made on top of the limited availability of data. The extraction of latent topics is mainly based on word co-occurrence, and, with only 94 documents, not only would the results be very limited, but there could not be any significant statistical analysis. This decision encompassed the change of approach from the traditional topic models to short-text topic models.

The models evaluated included the ones based on both internal and external semantic information. The extraction of internal semantic information is limited by the data availability. Results rendered this approach overfit to the documents, with almost imperceptible topics and poor aggregation of similar fragments in the projection space. The usage of external semantic information is limited by the domain adaptability and the collection's vocabulary. Results determined that, even though this approach showed better scores, it could not be taken to full advantage due to the limited richness of the vocabulary employed by the patients.

The semantic characterization, obtained with the analysis of the projection of the patients in the latent semantic space produced by the external semantic information short-text topic model, revealed the relative importance of the many aspects encompassing the experience of pain. Not only that, but it also reflected the engagement and outlook of each patient regarding the interview, and the various types of experiences of pain were identified and characterized. However, no relevant correlation was found between these types of experiences and demographic and clinical parameters. On the other hand, groups of patients given by these external parameters revealed that some groups report slightly different experiences, which is suggested to be related to the parameter itself.

### **7.1.3 Prediction of clinical parameters**

The prediction of clinical parameters presented in Chapter 6, based on the characterization obtained in previous experiments, revealed a specific application of the present study, in this case, the classification of pathology and pain intensity level based on verbal descriptions of pain. Even though the experimental setup only focused on these two parameters, the presented and discussed methodology may be applied to any parameter.

The best results obtained for pathology classification were based on vocabulary features, specifically utilizing the discussion of the aspect of the experience of pain related to the location on the body. These observations were found to be in line with the scientific research of the studied pathologies. The results obtained for pain intensity level classification were found

to be sub-optimal and various possible associated problems were identified, from the number of samples, to the class definition. Notably, all results were obtained in a Leave-One-Out validation setting due to the limited amount of samples. No result under this setting can be confidently generalized to a broader population.

## 7.2 *Future work*

In this section is proposed future work regarding both types of analysis, linguistic and paralinguistic. Most of the proposal stems from work that was intended to be performed, but could not be due to limited quality and availability of data. Thus, the following remarks expect a larger dataset, without sound and text quality limitations (or, at least, reduced to easily removable noise by the large number of samples).

The proposed future work regarding the linguistic analysis focuses on two aspects. First, an in-depth study of the population by question of the interview. Each question aims to discuss a specific aspect of the experience, thus, by understanding how each patient is positioned relative to others in each aspect (question), it would be possible to find relevant groups per aspect, and search for a more fine-grained correlation with external parameters. It was not possible to perform such an analysis with the current dataset, because the number of patients is very limited and the existing answers are too disperse. Second, the integration with the input provided by health professionals. This input includes the interpretation of health professionals regarding the clinical state of each patient solely based on the recording of each patient (there was no access to clinical or demographic parameters). Possible integration includes a similar topic modeling approach and a parallelism analysis between the computationally obtained results of the patients and the inputs provided by field professionals. This input could also help define ground truth labels to better evaluate the characterization analysis performed in Chapter 5.

Finally, regarding the paralinguistic analysis, almost all aspects were left undone due to the extremely poor audio quality. Emotion and speech disfluencies aspects were found to be relevant in the literature to the assessment and management of pain, and, thus, should be considered in future work. This includes the tasks of emotion recognition, sentiment analysis, and the identification of the various speech disfluencies, such as hesitations, repetitions, speed of speech, and others.



# Bibliography

- Azevedo, L. F., Costa-Pereira, A., Mendonça, L., Dias, C. C., & Castro-Lopes, J. M. (2012). Epidemiology of chronic pain: a population-based nationwide study on its prevalence, characteristics and associated disability in Portugal. *The Journal of Pain*, 13(8), 773–783.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901), 268–282.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Breivik, H., Borchgrevink, P., Allen, S., Rosseland, L., Romundstad, L., Breivik Hals, E., ... Stubhaug, A. (2008). Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1), 17–24.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4), 335.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).
- Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 163, 1–13.
- Chen, Z., & Liu, B. (2014). Topic modeling using topics from many domains, lifelong learning and big data. In *International Conference on Machine Learning* (pp. 703–711).
- Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 209–218).

- Cleeland, C., & Ryan, K. (1994). Pain assessment: global use of the Brief Pain Inventory. *Annals, Academy of Medicine, Singapore*.
- Dansie, E., & Turk, D. C. (2013). Assessment of patients with chronic pain. *British Journal of Anaesthesia, 111*(1), 19–25.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391–407.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dias, A. (2007). Dor Crónica–Um problema de saúde pública. *Psicologia*.
- Ehlich, K. (1985). The language of pain. *Theoretical Medicine, 6*(2), 177–187.
- Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459–1462).
- Fink, R. (2000). Pain assessment: the cornerstone to optimal pain management. In *Baylor university medical center proceedings* (Vol. 13, pp. 236–239).
- Gouveia, M., & Augusto, M. (2011). Custos indirectos da dor crónica em Portugal. *Revista Portuguesa de Saúde Pública, 29*(2), 100–107.
- Guzman, E., & Maalej, W. (2014). How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)* (pp. 153–162).
- Halliday, M. A. (1998). On the grammar of pain. *Functions of Language, 5*(1), 1–32.
- Hansen, G. R., & Streltzer, J. (2005). The psychology of pain. *Emergency Medicine Clinics, 23*(2), 339–348.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics), 28*(1), 100–108.

- Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., & Zimmermann, R. (2018). ICON: interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2594–2604).
- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296).
- Katz, J., & Melzack, R. (1992). Measurement of pain.
- Lascaratou, C. (2007). The language of pain. *Amsterdam and Philadelphia: John Benjamins*.
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788.
- Leino-Kilpi, H., Maenpaa, I., & Katajisto, J. (1999). Nursing study of the significance of rheumatoid arthritis as perceived by patients using the concept of empowerment. *Journal of Orthopaedic Nursing*, 3(3), 138–145.
- Lev, G., Klein, B., & Wolf, L. (2015). In defense of word embedding for generic text representation. In *International Conference on Applications of Natural Language to Information Systems* (pp. 35–50).
- Levene, H. (1960). Contributions to probability and statistics. *Essays in honor of Harold Hotelling*, 278–292.
- Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165–174).
- Loeser, J. D., & Melzack, R. (1999). Pain: an overview. *The lancet*, 353(9164), 1607–1609.
- Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.
- Luo, P. (2014 (accessed December 30, 2019)). *ACM IS abstract and citation network1*. Harvard Dataverse.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579–2605.

- Mamede, N., Baptista, J., Diniz, C., & Cabarrão, V. (2012). STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18–25).
- McVeigh, C. M., & Cairns, A. P. (2006). Diagnosis and management of ankylosing spondylitis. *Bmj*, 333(7568), 581–585.
- Melzack, R. (1975). The McGill Pain Questionnaire: major properties and scoring methods. *Pain*, 1(3), 277–299.
- Melzack, R. (2001). Pain and the neuromatrix in the brain. *Journal of dental education*, 65(12), 1378–1382.
- Melzack, R., & Torgerson, W. (1971). On the language of pain. *Anesthesiology*, 34(1), 50–59.
- Merskey, H., & Bogduk, N. (1994). Classification of chronic pain, IASP Task Force on Taxonomy. Seattle, WA: *International Association for the Study of Pain Press* (Also available online at [www.iasp-pain.org](http://www.iasp-pain.org)).
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.
- Miller, A., Green, M., & Robinson, D. (1983). Simple rule for calculating normal erythrocyte sedimentation rate. *British medical journal (Clinical research ed.)*, 286(6361), 266.
- Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2227–2231).
- Miyahara, K. (2019). Enactive pain and its sociocultural embeddedness. *Phenomenology and the Cognitive Sciences*, 1–16.

- Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3, 299–313.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pepys, M. B., Hirschfield, G. M., et al. (2003). C-reactive protein: a critical update. *The Journal of clinical investigation*, 111(12), 1805–1812.
- Pimenta, C., & Teixeira, M. (1996). Proposal to adapt the McGill Pain Questionnaire into Portuguese. *Revista da Escola de Enfermagem da USP*, 30(3), 473–483.
- Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 873–883).
- Qiang, J., Chen, P., Wang, T., & Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 363–374).
- Research, Y. (accessed December 30, 2019). Yahoo research datasets: Language data, 15 - yahoo! answers manner questions, version 2.0 [Computer software manual]. Retrieved from <https://webscope.sandbox.yahoo.com/catalog.php?datatype=1> (<https://webscope.sandbox.yahoo.com/catalog.php?datatype=1>)
- Reynolds, D. A. (2009). Gaussian Mixture Models. *Encyclopedia of biometrics*, 741.
- Rindfleisch, Adam J and Muller, Daniel. (2005). Diagnosis and management of rheumatoid arthritis. *American family physician*, 72(6), 1037–1047.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language*, 27(1), 4–39.

- Sebastian, J., & Pierucci, P. (2019). Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. *Proc. Interspeech 2019*, 51–55.
- Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1105–1114).
- Sridhar, V. K. R. (2019, March 26). *Unsupervised topic modeling for short texts*. Google Patents. (US Patent 10,241,995)
- Stein, C., & Mendl, G. (1988). The German counterpart to McGill pain questionnaire. *Pain*, 32(2), 251–255.
- Sullivan, M. D. (1995). Pain in language: from sentience to sapience. In *Pain Forum* (Vol. 4, pp. 3–14).
- Sussex, R. (2009). Review article of Chryssoula Lascaratou's the language of pain. *Australian Review of Applied Linguistics*, 32(1), 6–1.
- Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., ... Gonçalves, M. A. (2019). Clu- Words: Exploiting Semantic Word Clustering Representation for Enhanced Topic Mod- eling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 753–761).
- Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S., Mourão, F., ... Gonçalves, M. A. (2018). Semantically-Enhanced Topic Modeling. In *Proceedings of the 27th ACM International Con- ference on Information and Knowledge Management* (pp. 893–902).
- Vorontsov, K., & Potapenko, A. (2015). Additive regularization of topic models. *Machine Learning*, 101(1-3), 303–323.
- Wilson, D., Williams, M., & Butler, D. (2009). Language and the pain experience. *Physiotherapy Research International*, 14(1), 56–65.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37–52.
- Wolf, C. J. (2010). What is this thing called pain? *The Journal of clinical investigation*, 120(11), 3742–3744.

- Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456).
- Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233–242).
- Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.
- Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2105–2114).





# I Appendices



# A

## Patient Profiling

In this appendix is made an exposition of the distribution of the population according to select demographic and clinical parameters, as well as linguistic and paralinguistic features. This is presented to help contextualize the core of the work. This is followed by the proposition of a tool to automatically produce a patient profile which merges the clinical and linguistic domains to aid with the assessment and management of the patient.

### A.1 *Population distribution*

In this section we present a report on the distribution of the population under study. This includes the demographic and clinical distributions, as well as the distribution of parameters in the linguistic and paralinguistic domains, per question, and per interviewer.

#### A.1.1 **Demographic distribution**

Female	Male	Other		Total
61	33	0		94
Active	Medical leave / Retired	Not specified		Total
46	39	9		94
Primary	High school	Bachelor	Master	Total
64	19	7	3	94

Table A.1: Number of subjects per demographic parameter.

We are interested in understanding the demographic distribution of the population under study, namely features such as age, level of studies and professional activity (Figure A.1). We observe that most subjects are on the older half of the spectrum (above 50 years of age). This is in line with the studied correlation of advanced age and chronic pain disease. Roughly 65% of the subjects belong to the feminine gender, with an age distribution varying all across the spectrum. Again, this is in line with the studied correlation of the feminine gender and

chronic pain disease. The majority of the population (roughly 68%) holds the primary level of education. This group also accommodates for a wide distribution on the age spectrum. The remaining education level groups are too small to hold any representational information. Finally, regarding professional activity, we observe that there seems to be a similar distribution of subjects still active and subjects under medical leave or retired. Those under medical leave or retired include the oldest part of the population, as expected.

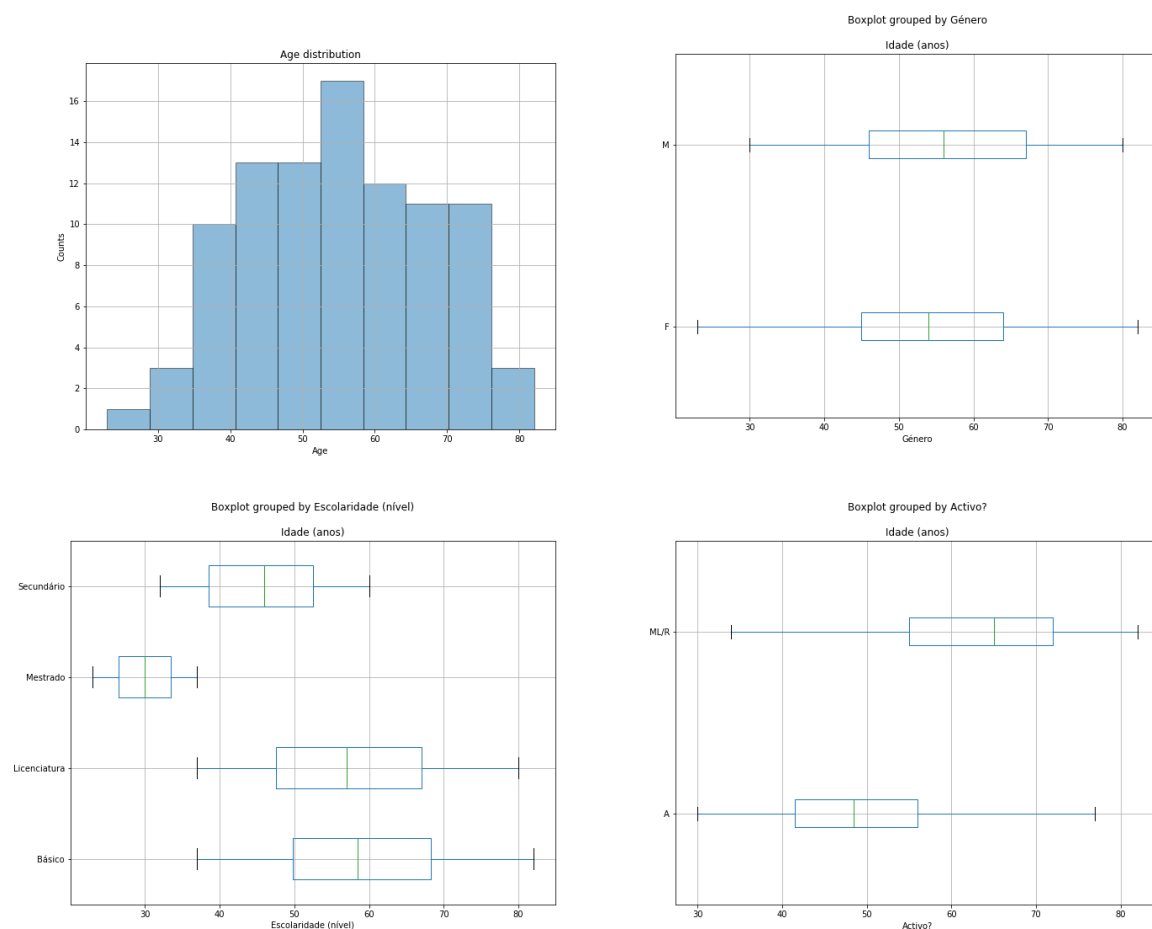


Figure A.1: Demographic parameters distribution over the age spectrum.

### A.1.2 Clinical distribution

Rheumatoid arthritis	Spondyloarthritis	Other
41	45	8

Table A.2: Number of subjects per identified pathology.

We are interested in understanding the clinical distribution of the population under study, namely features such as pathology, pain intensity, duration of the disease, ESR (VS) and RCP (PCR) (Figure A.2). It is also relevant to understand the variation of these features given the age group (Figure A.3). The population is similarly distributed between the pathologies of Rheumatoid Arthritis and Spondyloarthritis. Self-reported pain intensity (on a visual analogue scale) ranges from 0 to 100, with the maximum reported intensity being 94. This parameter's boxplot highlights that the subject's reports are evenly distributed across the spectrum, with a slight tendency to lower scores (less pain). Looking at pain intensity distribution across age bins, the group that reports the highest pain intensities is the middle tear age group. A study against the professional activity (omitted) did not show a significant correlation between the professionally active group and reported pain. The duration of the disease expectedly fits with the group's age distribution.

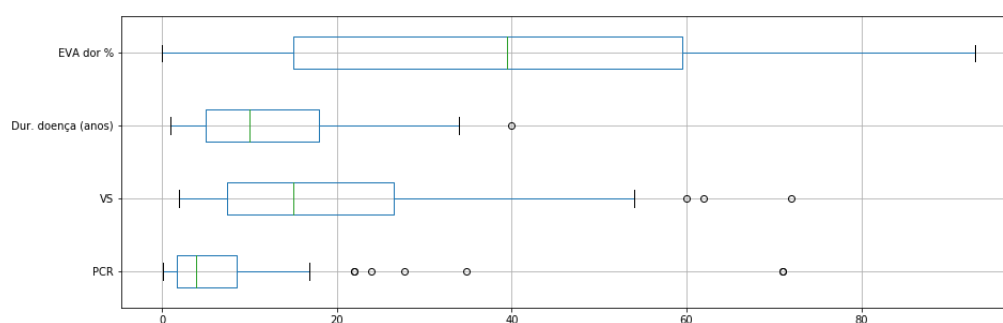


Figure A.2: Distribution of the clinical parameters.

Regarding the correlation between these parameters (Table A.3), it is not possible to observe any with pain intensity, although there are some slight suggestions, especially with age (older patients report higher pain intensities). The only notable observation at this point is the somewhat suggestive correlation of VS and PCR.

	Idade (anos)	EVA dor %	Dur. doença (anos)	VS	PCR
Idade (anos)	1.000	0.200	0.267	0.323	-0.030
EVA dor %	0.201	1.000	0.125	0.143	0.025109
Dur. doença (anos)	0.267	0.125	1.000	-0.069	-0.014
VS	0.323	0.143	-0.069	1.000	0.341687
PCR	-0.031	0.025	-0.014	0.342	1.000

Table A.3: Correlation between select clinical and demographic parameters.

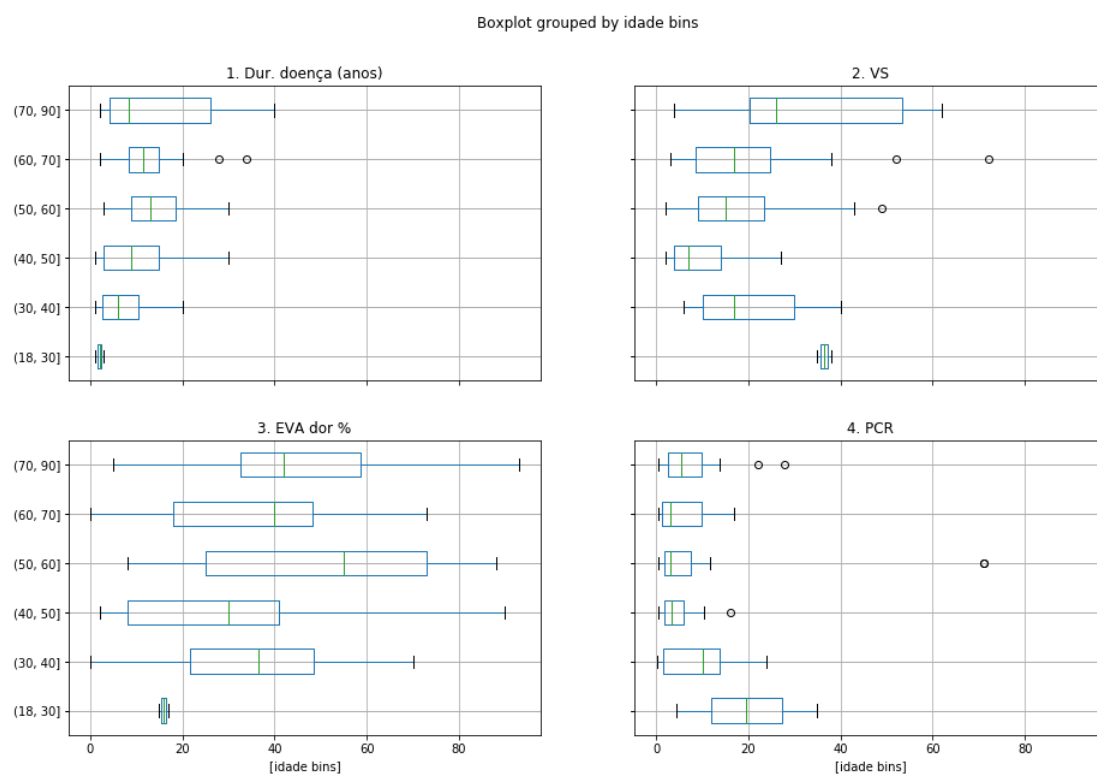


Figure A.3: Distribution of the clinical parameters given age bins.

Analyzing these parameters by interviewer (Figure A.4), we observe that the sub-populations share a similar distribution regarding the duration of the disease and the PCR values. However, indeed the patients' data collected by SP report less intense pain, whilst CV and DO have very similar distributions on this dimension.

### A.1.3 Linguistic and paralinguistic analysis

As it can be assessed from the interview script, the nature of each question (and consequently the respective answer) is different. Some questions require a higher development and engagement from the patient, than others, to which a direct answer usually suffices. This requires the analysis to be both inter- and intra-questions. The following sections explore the recordings and transcriptions through these lenses.

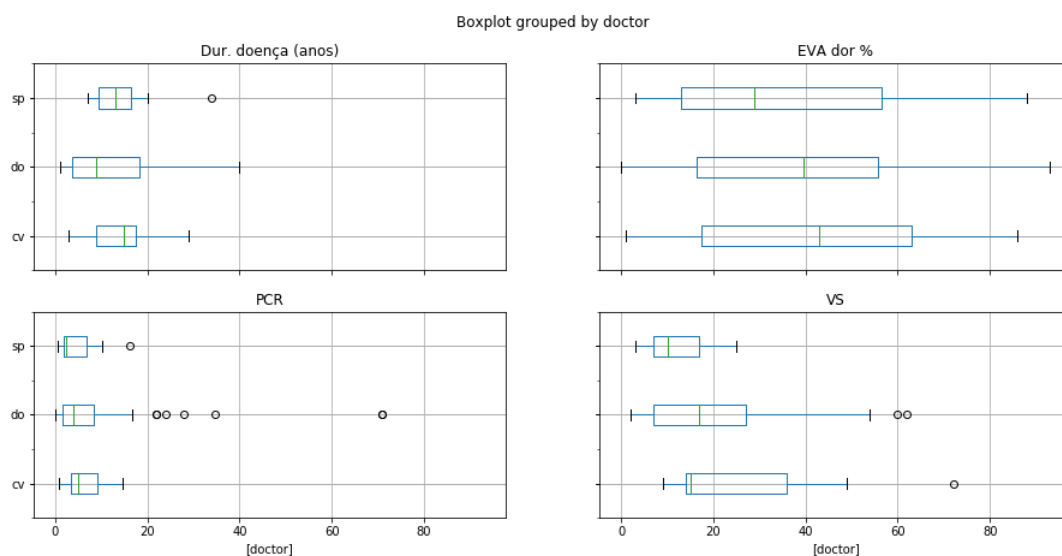


Figure A.4: Distribution of the clinical parameters by interviewer.

### A.1.3.1 Inter-question

Comparing questions regarding word counts (Figure A.5), indeed it is observed that specific questions elicit more words than others. Specifically, Q4 leads with the maximum word counts. This observation fits with the interview design expectations. Conversely, Q1 and Q5 have the least word counts, with Q5 being slightly lower. Given that Q5 is a complex question, and it is comparable to Q1 on this remark, which is a direct, closed answer, we conclude that it did not elicit the expected engagement from the patients. Taking into account the size of the population, the number of outliers cannot be disregarded.

Under the same question aggregation, textual information richness is measured by the TF-IDF score. Expectedly, the more words present in a set of documents, the more likely it is to have a higher cumulative TF-IDF score. Q4, again, leads with maximum values, with Q6 in a close second maximum. Q1 and Q5 carry the least rich textual information. Q3 and Q6 have very similar distributions, which coincides with the very similar nature of both questions.

Finally, looking at the word rate, by question (number of words per second of recorded answer), we conclude that there is indeed some expected variability, but the distributions between questions are very similar.

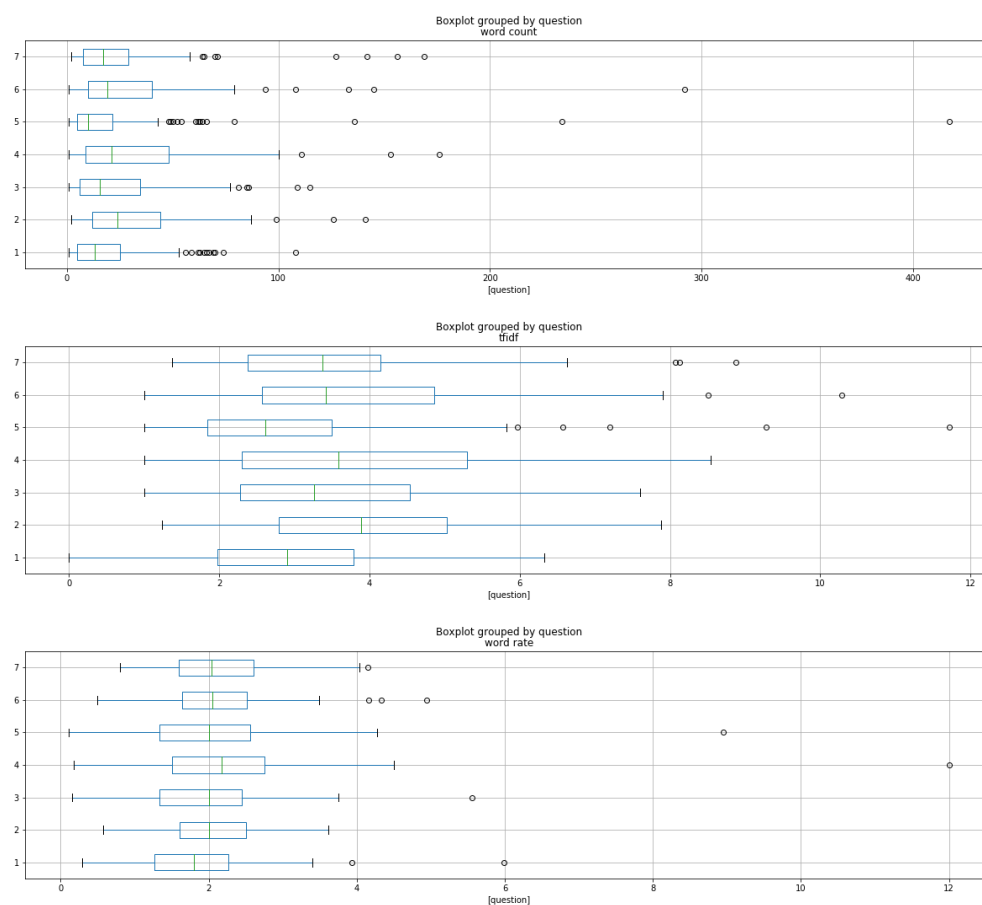


Figure A.5: In order, word count, TF-IDF, and word rate features distributions across questions.

### A.1.3.2 Interviewer engagement

In an attempt to reflect the level of engagement the interviewer had with the patients, we look at the distributions of text features aggregated by this parameter (Figure A.6). In a straightforward observation, we conclude that, on average, CV elicits more engagement than any other interviewer. The inverse is observed for SP.

### A.1.3.3 Intra-question (by interviewer)

Regarding word counts (Figure A.7), we conclude that for most questions, the interviewer does not have an impactful bias in the distributions (attending to the fact that SP questions are constantly a little behind). However, Q4 and Q6 both show a lot of disparity given the interviewer. The same observation can be made for the word rate feature (Figure A.8). TF-IDF score, on the other hand, shows greater disparity in most questions' distributions (Figure A.9).



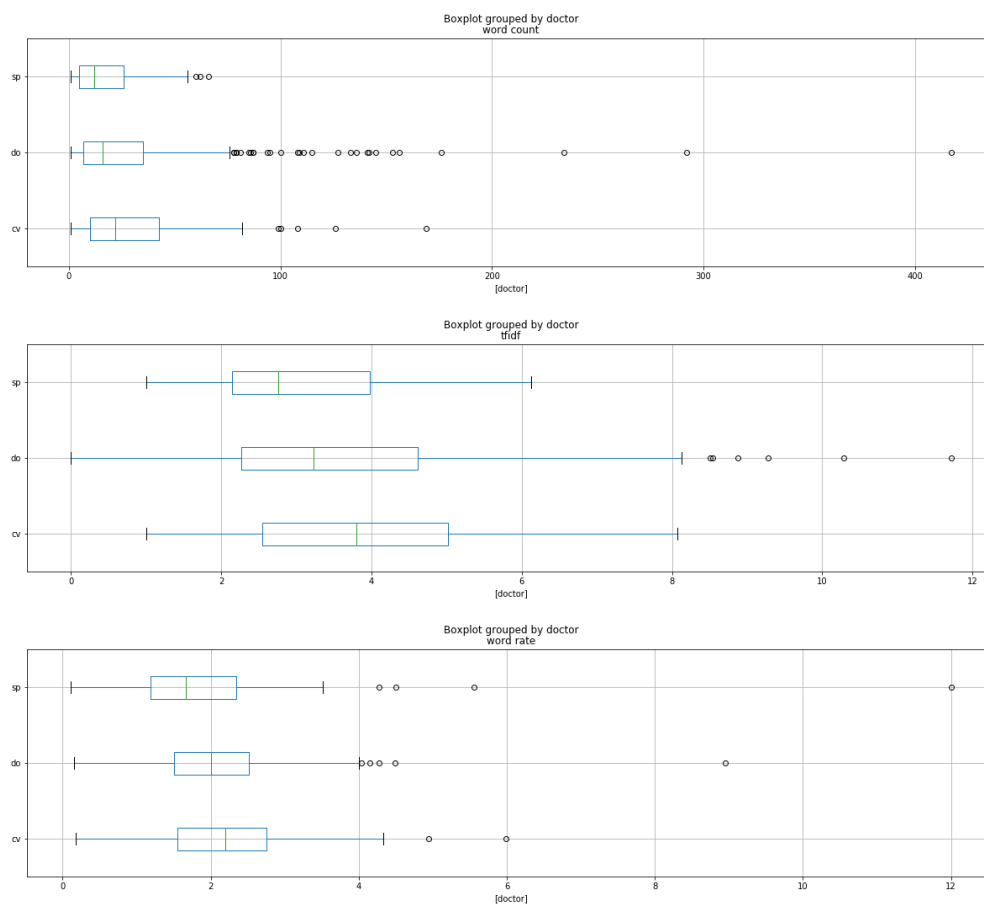


Figure A.6: In order, cumulative word count, TF-IDF, and word rate features distributions across interviewers.

#### A.1.3.4 Correlation with demographic and clinical features

An analysis of linguistic and paralinguistic features against demographic and clinical features shows no concrete evidence of correlation between these domains (Table A.4).

	word count	TFIDF	word rate	audio length
Idade (anos)	-0.111	-0.122	-0.090	-0.089569
EVA dor %	0.090	0.123	0.118	0.074
Dur. doença (anos)	-0.040	-0.059	-0.044	-0.033
VS	0.011	0.017	0.035	-0.002
PCR	0.005	-0.001	-0.018	0.008

Table A.4: Correlation between clinical and demographic parameters with linguistic features.

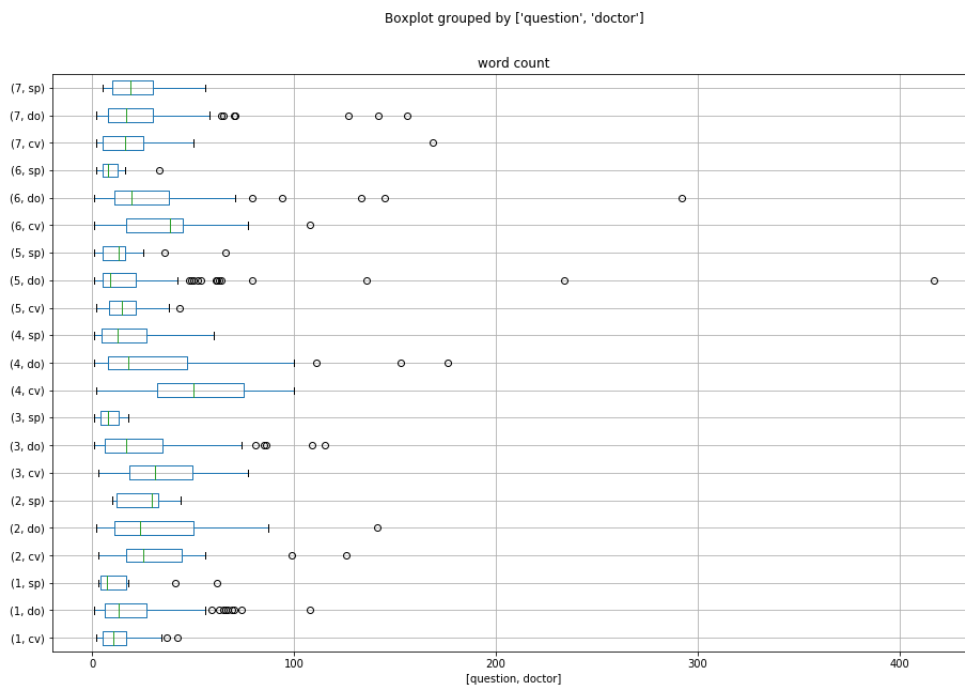


Figure A.7: Word count feature distribution by question and by interviewer.

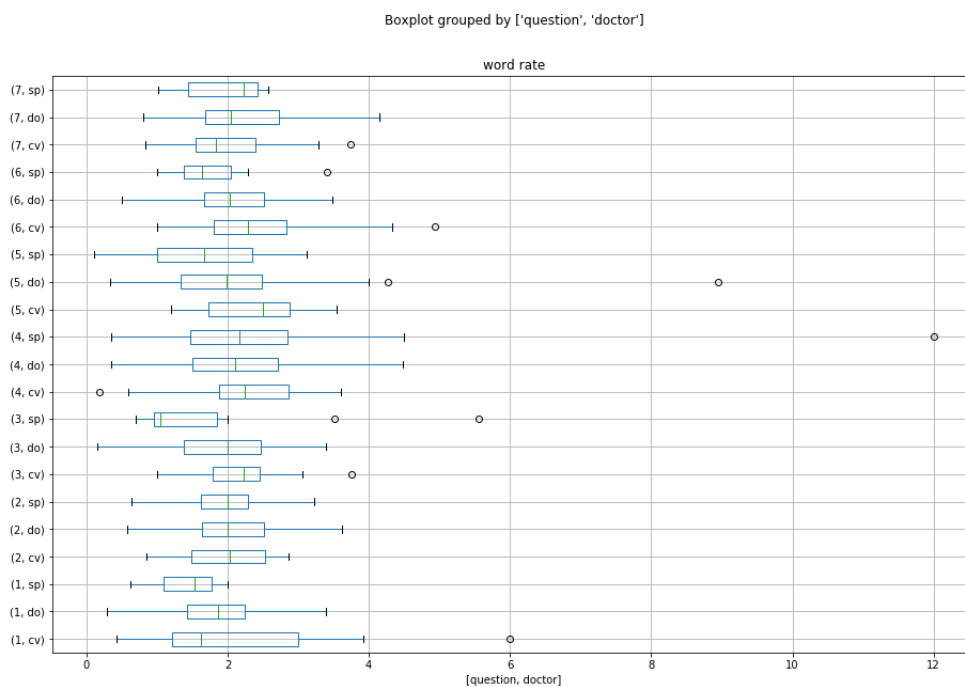


Figure A.8: Word rate feature distribution by question and by interviewer.

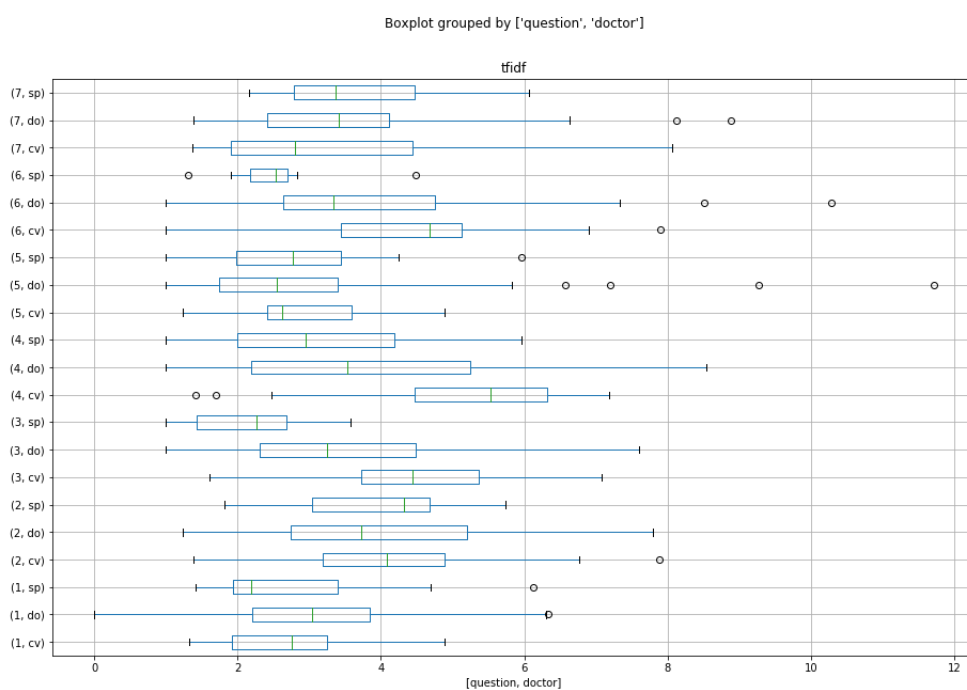


Figure A.9: TF-IDF feature distribution by question and by interviewer.

## A.2 Profiling tool

An experience of pain may be expressed through a multitude of channels, both in objective and subjective forms. The clinical parameters, which are by themselves a manifestation of the clinical situation of the patient, provide an objective, but limited, perspective on the development of the disease and pain. On the other hand, subjective forms of pain manifestations, such as a verbal expression, which encompasses both semantics and paralinguistics, provide useful, but complex information, because they are, by definition, influenced by multiple layers of subjectivity. Only when considering the multiple dimensions in which pain and the disease manifest, can the health professional reliably provide a quality assessment.

In this appendix we propose a methodology to bring together multiple dimensions of pain, specifically textual descriptions of pain and clinical parameters, in a condensed, profiling format, so that the health professional can have access to a global view of each patient regarding their pain and disease.

The patient profile is defined by two major parts, the linguistic analysis of their interview, and the analysis of their corresponding clinical form, denominated the clinical panel. The linguistic analysis includes the distribution of semantic topics, determined by the whole collection of documents, which were found to be the most relevant for the specific patient (as described in the previous chapter), as well as the identification of the most important words. Additionally, pain locations on the body are extracted from the text and represented on a human body diagram. The clinical panel of a patient includes the values of pain duration, Visual Analogue Scale (VAS) of both pain and disease, Erythrocyte Sedimentation Rate (ESR) and C-reactive protein (CRP). The following sections describe each of these values, and how they can be represented in a meaningful way for health professionals.

### A.2.1 Linguistic domain

The linguistic analysis of a patient's description of their personal experience of pain includes the 3 following components.

For the first component, the latent topics are extracted from the interview with each patient, as described in Chapter 5. Each of the topics (restricted to the top 10 most weighted words) is interpreted and labeled for ease of read. The importance given by the patient to each of these

topics is displayed in a radar plot, so that patients with similar topic importances display a similar map.

For the second component, a wordcloud is built, such that the words expressed by the patient that are considered to be the most relevant given a specific metric, are highlighted by their size. This metric is defined to be the TF-IDF metric, because it allows for a positioning of the patient's words in relation to the whole collection. Observing this kind of wordcloud, one can quickly interpret the main preoccupations of that patient.

Finally, for the third component, the words in the collection's vocabulary that relate to locations of the human body are identified (e.g. knee, leg, hand, wrist, etc.) and the corresponding frequency values are extracted. With the help of a human body diagram, markers are placed in their respective locations. The size of the markers reflect the frequency values, or, in other words, the importance of those locations to the patient.

### A.2.2 Clinical domain

As stated before, the patient clinical panel includes all of the pain, disease, and clinical parameters provided by the complementary form. Additionally, it includes a relative value, in a 0-100 scale, obtained from the mean of all other values and scaled according to all other patients to obtain a comparative scale. The patient which scores 100 in this comparative scale is the one whose mean value of all other parameters is the highest in the group (which suggests that it is, in general, the patient in worse condition, from those included in the study).

All values presented in the clinical panel are accompanied by a color scale to facilitate their assessment. This color scale is white for neutral values, and a gradient from yellow to red (best to worst), for values incrementally further from the values considered neutral. The VAS of both pain and disease range from 0 to 100. Zero is considered to be the neutral value. The duration of symptoms of pain can range from 0 to an unknown (a priori) upper limit. Regarding the ESR parameter, the widely used rule to calculate the normal maximum ESR value (Miller, Green, & Robinson, 1983) was used to determine, for each patient, the maximum value considered neutral. Finally, regarding the CRP parameter, in healthy adults, the normal values range between 0.8 mg/L and 3.0 mg/L (Pepys, Hirschfield, et al., 2003), and, thus, values increasingly farther from the 3.0 mark are considered increasingly worse in the color

scale. The upper limit of these features is defined as the highest value in the sample of patients, for each feature.

### A.2.3 Results

For the first component of the linguistic domain, from the 7 documents of each patient (corresponding to the 7 questions in the interview) were extracted 10 semantic topics, converted into vector form and aggregated into a single vector, obtained by the mean value of each dimension (topic). The topic model used was CluWords with FastText. Each of the topics (restricted to the top 10 most weighted words) was interpreted and labeled for ease of read. The labels are as follows: (1) expectations, (2) pain location, (3) medication, (4) pain evolution, (5) body sensations, (6) effects, (7) reflections, (8) sources of pain, (9) pain location, (10) evolution/medication. The resulting radar plot can be observed in Figure A.10a. For the second component, using the intermediate vector representation of the collection obtained with TF-IDF (CluWords), a wordcloud is built, such that the words expressed by the patient, considered to be the most relevant given the TF-IDF metric, are highlighted by their size. Looking at this kind of wordcloud, one can quickly identify the main preoccupations of the patient. Such a wordcloud can be observed in Figure A.10b. For the third component of the linguistic domain, the words in the collection's vocabulary that relate to locations of the human body were identified and with the help of a human body diagram, markers were placed in their respective locations. The size of the markers reflects the TF-IDF (CluWords) values (which is also present in the previously discussed wordcloud). An example of this diagram can be observed in Figure A.10c. Finally, an example of a clinical panel with color gradients over the clinical parameters can be observed in Figure A.11. All of these components are displayed side-by-side, in an easy to interpret profile, as exemplified in Figure A.12.



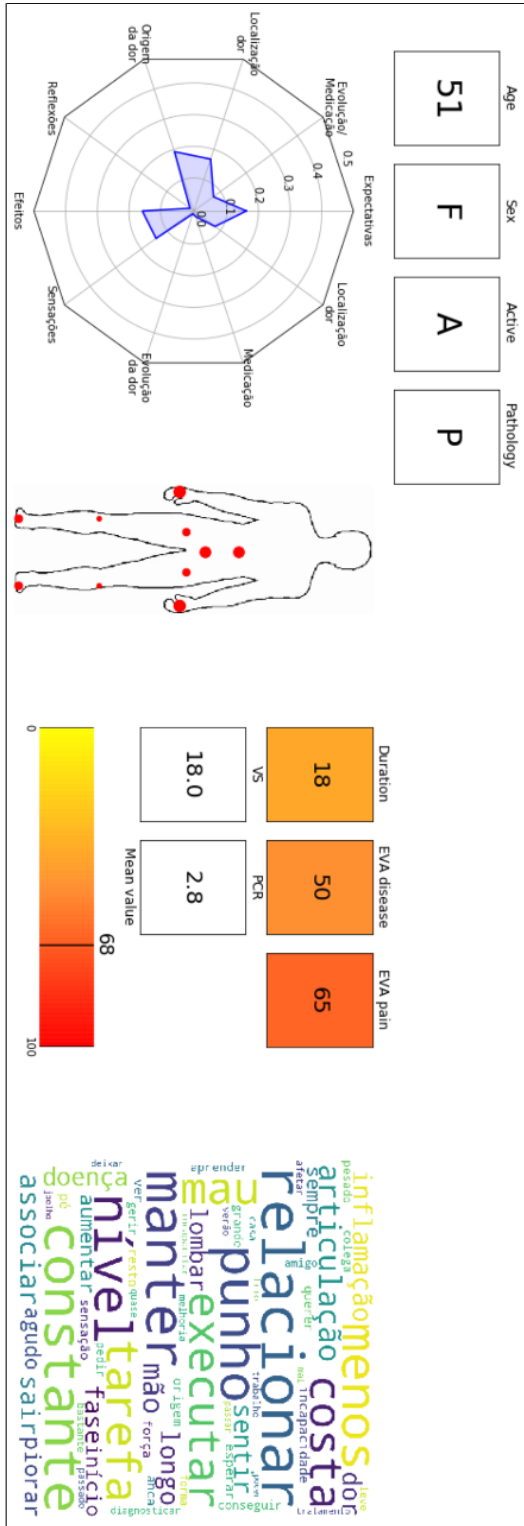


Figure A.12: Example of a complete patient profile.



# B

## Figures and Tables

In this appendix we present some of the results that were omitted from the main discussion, as well as complementary tables with the actual values corresponding to these plotted results.

### B.1 Pathology classification

	fragment	patient	sq 1	sq 2	sq 3
EXP 1	$0.53 \pm 0.003$	$0.54 \pm 0.003$	$0.54 \pm 0.005$	$0.51 \pm 0.005$	$0.48 \pm 0.004$
EXP 2	$0.55 \pm 0.002$	$0.56 \pm 0.004$	$0.64 \pm 0.003$	$0.43 \pm 0.011$	$0.46 \pm 0.010$
EXP 3	$0.62 \pm 0.005$	$0.57 \pm 0.002$	$0.69 \pm 0.004$	$0.52 \pm 0.001$	$0.51 \pm 0.006$
EXP 4	$0.64 \pm 0.005$	$0.60 \pm 0.005$	$0.71 \pm 0.003$	$0.53 \pm 0.004$	$0.48 \pm 0.009$
	sq 4	sq 5	sq 6	sq 7	
EXP 1	$0.45 \pm 0.002$	$0.52 \pm 0.002$	$0.47 \pm 0.002$	$0.44 \pm 0.010$	
EXP 2	$0.49 \pm 0.002$	$0.54 \pm 0.001$	$0.52 \pm 0.002$	$0.47 \pm 0.001$	
EXP 3	$0.44 \pm 0.001$	$0.62 \pm 0.004$	$0.48 \pm 0.002$	$0.48 \pm 0.016$	
EXP 4	$0.48 \pm 0.004$	$0.56 \pm 0.007$	$0.51 \pm 0.008$	$0.52 \pm 0.007$	

Table B.1: Mean accuracy score of each experiment in Table 6.6, over the different types of feature aggregation in Table 6.4. Table of values in Figure 6.1.

feature	fragment	patient	sq 1	sq 5
BoW	0.63	0.57	0.72	0.60
tf-idf	0.67	0.66	0.78	0.63
LDA	0.51	0.56	0.60	0.60
NMF	0.72	0.56	0.72	0.71

(a) Experiment 3.

feature	fragment	patient	sq 1	sq 5
BoW	0.67	0.62	0.77	0.57
tf-idf	0.67	0.64	0.79	0.53
LDA	0.51	0.49	0.69	0.36
NMF	0.66	0.71	0.62	0.64

(b) Experiment 4.

Table B.2: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Focused only on the baselines. Table of values in Figure 6.2.

feature	fragment	patient	sq 1	sq 5
tf-idf	0.67	0.66	0.78	0.63
NMF	0.72	0.56	0.72	0.71
SeaNMF	0.57	0.53	0.66	0.50
CluWords (FastText)	0.64	0.60	0.67	0.61
CluWords (BERT)	0.65	0.53	0.72	0.68
BERT	0.55	0.51	0.60	0.65

(a) Experiment 3.

feature	fragment	patient	sq 1	sq 5
tf-idf	0.67	0.64	0.79	0.53
NMF	0.66	0.71	0.62	0.64
SeaNMF	0.69	0.64	0.71	0.58
CluWords (FastText)	0.70	0.60	0.70	0.58
CluWords (BERT)	0.63	0.60	0.71	0.60
BERT	0.56	0.51	0.71	0.58

(b) Experiment 4.

Table B.3: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Some baseline results were omitted for ease of read. Table of values in Figure 6.3.

## B.2 Pain intensity classification

	fragment	patient	sq 1	sq 2	sq 3
EXP 1	0.34 ± 0.006	0.33 ± 0.006	0.32 ± 0.003	0.31 ± 0.006	0.37 ± 0.001
EXP 2	0.35 ± 0.006	0.33 ± 0.002	0.24 ± 0.004	0.30 ± 0.009	0.35 ± 0.003
EXP 3	0.34 ± 0.006	0.35 ± 0.000	0.27 ± 0.001	0.30 ± 0.001	0.42 ± 0.001
EXP 4	0.36 ± 0.003	0.35 ± 0.001	0.32 ± 0.001	0.33 ± 0.002	0.29 ± 0.003
	sq 4	sq 5	sq 6	sq 7	
EXP 1	0.36 ± 0.003	0.27 ± 0.002	0.32 ± 0.002	0.32 ± 0.003	
EXP 2	0.36 ± 0.007	0.29 ± 0.005	0.28 ± 0.002	0.30 ± 0.003	
EXP 3	0.35 ± 0.007	0.33 ± 0.002	0.34 ± 0.001	0.38 ± 0.001	
EXP 4	0.33 ± 0.003	0.37 ± 0.004	0.33 ± 0.003	0.42 ± 0.006	

Table B.4: Mean accuracy score of each experiment in Table 6.6, over the different types of feature aggregation in Table 6.4. Table of values in Figure B.1.

feature	fragment	patient	sq 1	sq 2	sq 3
BoW	0.36	0.34	0.29	0.30	0.42
tf-idf	0.47	0.37	0.29	0.30	0.46
LDA	0.31	0.37	0.23	0.31	0.38
NMF	0.41	0.33	0.28	0.24	0.43
feature	sq 4	sq 5	sq 6	sq 7	
BoW	0.44	0.37	0.34	0.38	
tf-idf	0.50	0.38	0.36	0.44	
LDA	0.30	0.37	0.40	0.36	
NMF	0.33	0.33	0.30	0.38	

(a) Experiment 3.

feature	fragment	patient	sq 1	sq 2	sq 3
BoW	0.34	0.41	0.33	0.32	0.25
tf-idf	0.42	0.34	0.35	0.36	0.28
LDA	0.38	0.35	0.31	0.36	0.22
NMF	0.34	0.37	0.36	0.27	0.29
feature	sq 4	sq 5	sq 6	sq 7	
BoW	0.36	0.36	0.37	0.44	
tf-idf	0.37	0.33	0.38	0.44	
LDA	0.32	0.31	0.38	0.52	
NMF	0.37	0.40	0.27	0.48	

(b) Experiment 4.

Table B.5: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Focused only on the baselines. Table of values in Figure B.2.

feature	fragment	patient	sq 1	sq 2	sq 3
tf-idf	0.47	0.37	0.29	0.30	0.46
LDA	0.31	0.37	0.23	0.31	0.38
NMF	0.41	0.33	0.28	0.24	0.43
SeaNMF	0.24	0.36	0.26	0.28	0.36
CluWords (FastText)	0.34	0.35	0.30	0.28	0.40
CluWords (BERT)	0.38	0.31	0.26	0.34	0.46
BERT	0.24	0.35	0.26	0.33	0.40
feature	sq 4	sq 5	sq 6	sq 7	
tf-idf	0.50	0.38	0.36	0.44	
LDA	0.30	0.37	0.40	0.36	
NMF	0.33	0.33	0.30	0.38	
SeaNMF	0.31	0.31	0.36	0.34	
CluWords (FastText)	0.24	0.26	0.33	0.41	
CluWords (BERT)	0.35	0.29	0.28	0.42	
BERT	0.31	0.33	0.36	0.34	

(a) Experiment 3.

feature	fragment	patient	sq 1	sq 2	sq 3
tf-idf	0.42	0.34	0.35	0.36	0.28
LDA	0.38	0.35	0.31	0.36	0.22
NMF	0.34	0.37	0.36	0.27	0.29
SeaNMF	0.26	0.29	0.33	0.32	0.38
CluWords (FastText)	0.34	0.36	0.26	0.33	0.35
CluWords (BERT)	0.42	0.33	0.29	0.38	0.28
BERT	0.37	0.34	0.34	0.27	0.29
feature	sq 4	sq 5	sq 6	sq 7	
tf-idf	0.37	0.33	0.38	0.44	
LDA	0.32	0.31	0.38	0.52	
NMF	0.37	0.40	0.27	0.48	
SeaNMF	0.40	0.29	0.28	0.35	
CluWords (FastText)	0.33	0.44	0.31	0.41	
CluWords (BERT)	0.27	0.46	0.27	0.46	
BERT	0.23	0.35	0.39	0.28	

(b) Experiment 4.

Table B.6: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Some baseline results were omitted for ease of read. Table of values in Figure B.3.

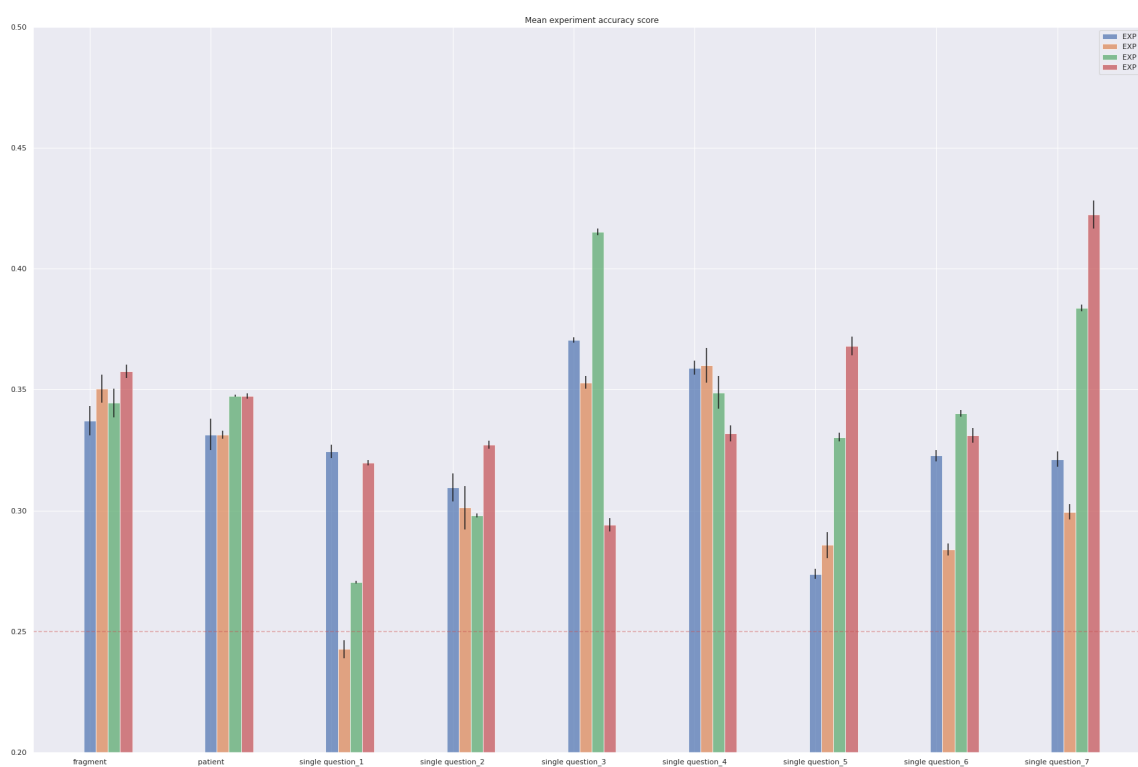
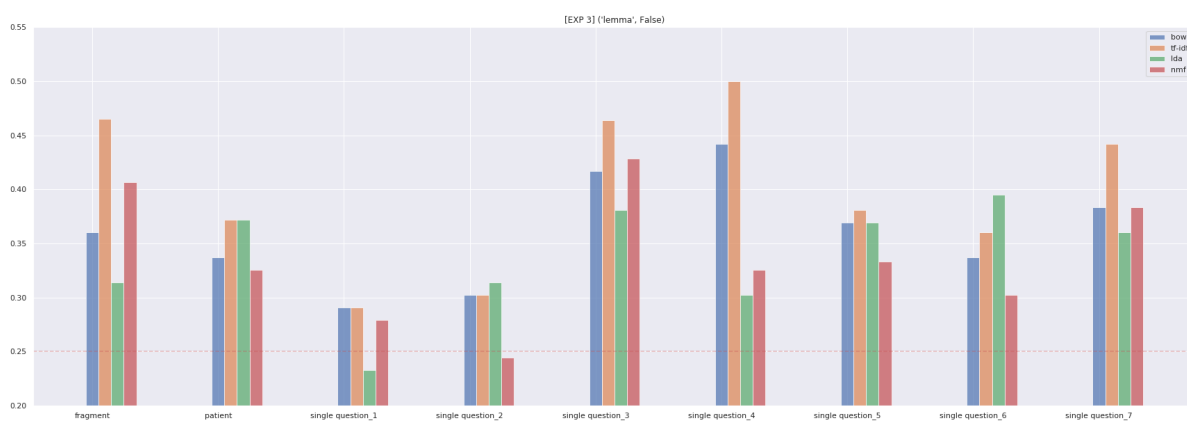
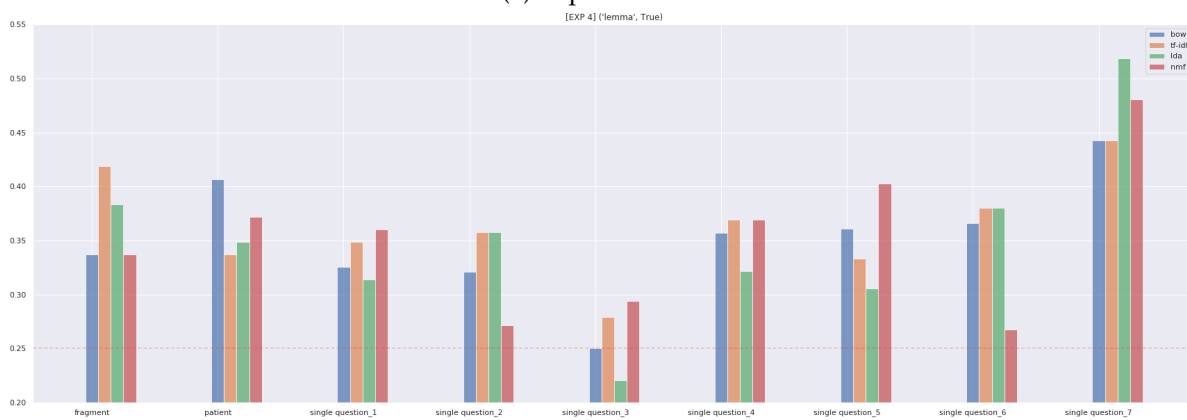


Figure B.1: Mean accuracy score of each experiment in Table 6.6, over the different types of feature aggregation in Table 6.4.

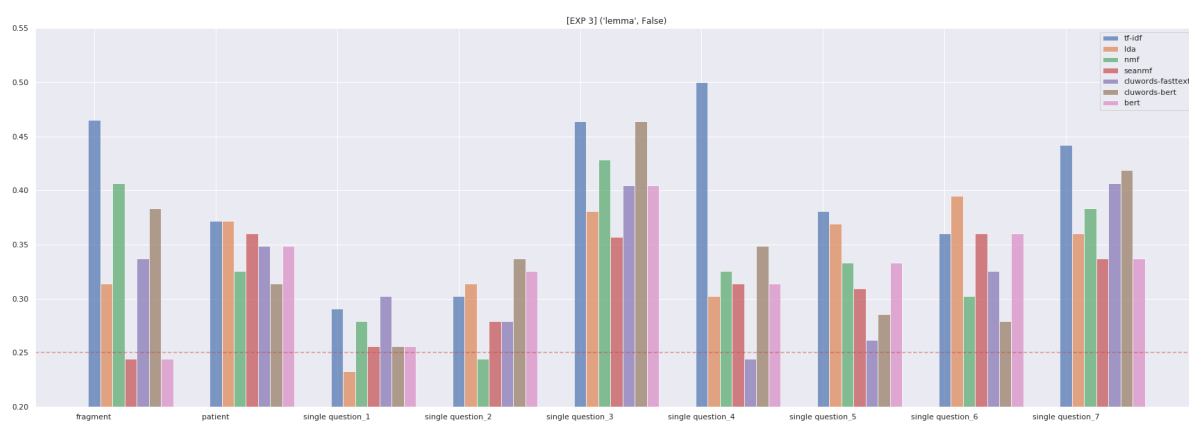


(a) Experiment 3.

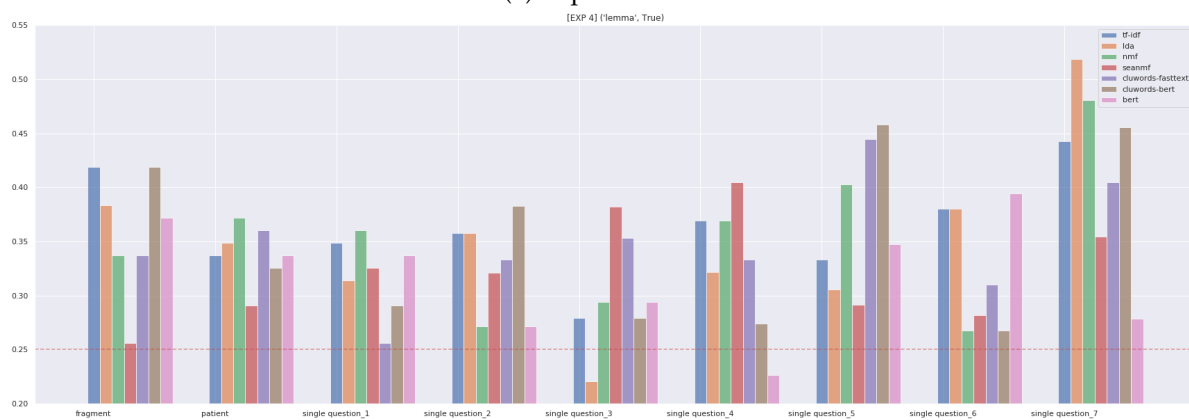


(b) Experiment 4.

Figure B.2: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Focused only on the baselines.



(a) Experiment 3.



(b) Experiment 4.

Figure B.3: Accuracy score of each feature type in Table 6.3, over the different types of feature aggregation in Table 6.4. Some baseline results were omitted for ease of read.

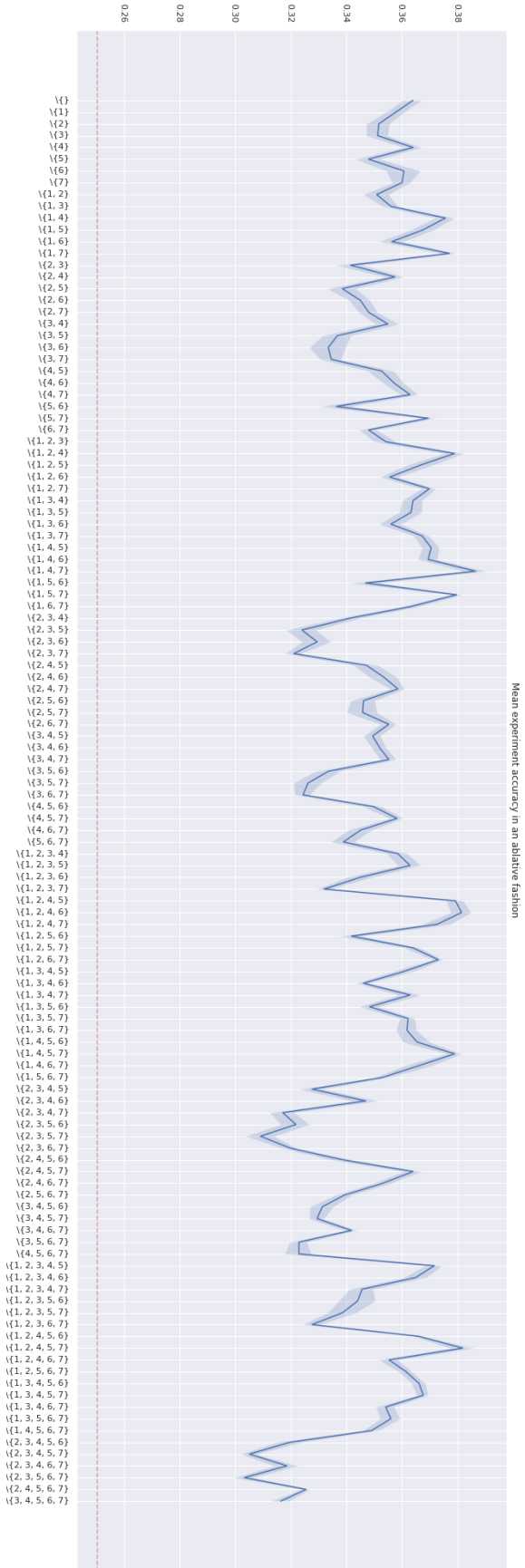


Figure B.4: Mean accuracy score of all experiments in Table 6.6, for all feature types in Table 4.2, in an ablativ fashion. The horizontal label indicates which set of questions is removed from the dataset prior to training and evaluation.