Chronic Pain Assessment from Patient Reports

Diogo Nunes
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
diogo.p.nunes@tecnico.ulisboa.pt

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 physiological 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 $\sim 80\%$. Conversely, 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.

Keywords: Chronic Pain, Pain Perception, Computational Pain Assessment, Topic Models for Pain, Information Extraction from Speech

1. Introduction

Pain is a subjective and private experience. It is subjective because it is dependent on biomedical, psychological, and sociocultural dimensions that directly influence how it is perceived and consequently expressed by the subject in pain. These encompass the patient's perception matrix. 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, verbal descriptions, to changes in behavior. These, together with demographic and clinical parameters related to the physiological manifestation of pain, are the inputs used by health professionals to assess and manage pain.

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.

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 subject perception and the underlying pain mechanisms, which are relevant details for an adequate pain management. Thus, the analysis of the language of pain, specifically trough 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 is that semantically related descriptions of pain may represent related experiences and can indirectly characterize the different types of pain. Concretely, the objective of the present work is twofold, given a population of verbal 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.

The document is structured as follows. Section 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. Section 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 methods. Section 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. Finally, Sections 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 clinical parameters.

2. The Nature of Pain

Pain is a sensation and an experience that issues a warning that something is probably wrong with the body. 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 twofold: 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, effectively 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.

A chronic pain experience is characterized by its persistent state, either continuous or recurrent, 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 and 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 and 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 and 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 and 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 and Melzack, 1999; Fink, 2000; Breivik et al., 2008; Azevedo et al., 2012).

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 and 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 this suffering 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 and 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 pain experience together with regions of the brain involved in affective and cognitive activities (Loeser and Melzack, 1999). In essence, sensory inputs are fed into the neuromatrix which 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.

2.2. *Linguistic expression of pain*

The experience of pain is only accessible to the outside world through an outward expression or behavior, rendering this 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 and 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 and 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 et al., 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.

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.

3. Pain and Language Analysis

Pain assessment is the cornerstone for its management. An adequate assessment will provide significant insights to the extent and magnitude of the disease, and the development of the recovery process. A linguistic analysis of the patients' description of pain may provide insights on the aforementioned relevant factors to the assessment. Specifically, it has been stated that similar descriptions of pain might describe similar characteristics of different experiences of pain. Allowing these descriptions to be characterized by their semantic topics allows us 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.

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. *Short-text topic modeling*

Topic modeling 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 thought of 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 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. 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 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 do not have the restrictions identified in short-text documents. 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 $\mathbf{w} = (w_1 w_2 ... w_N)$; and a collection of M documents is denoted $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

Topic modeling methods follow either probabilistic or non-probabilistic approaches. Non-probabilistic approaches, such as the non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999) model, 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 $\mathrm{TFIDF}_{|V|\times M}$ matrix. On the other hand, probabilistic approaches, such as the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model, assume a generative probabilistic process for each document. With short texts, probabilistic approach-based models have been shown to under-perform when compared against non-probabilistic-based models (Chen et al., 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.

The semantics-assisted NMF (SeaNMF) (Shi et al., 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 (Fig. 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 $S \approx W W_c^T$. 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. (1), 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. It is argued that the fact that the semantic information is learned from the collection itself, and not from an external source, is a determining factor due to the possibility of introducing bias from context-inadequate semantic spaces.

$$\min_{W,W_C,H\geq 0} \left\| \begin{bmatrix} N^T \\ \sqrt{\alpha}S^T \end{bmatrix} - \begin{bmatrix} H \\ \sqrt{\alpha}W_c \end{bmatrix} W^T \right\|_F^2 + \psi(W, W_c, H) \tag{1}$$

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 α 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. (2). In this approach, matrix $\mathrm{C}_{\mathrm{tf}M\times|V|}$ represents the term frequencies of each CluWord in each document, so that row $\mathrm{C}_{\mathrm{tf}d}$ 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). Matrix $\mathrm{idf}(C)$ determines the inverse document frequency of each CluWord

$C_t \in C$ as defined in Eq. (4). 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_{tfidf} is then submitted to factorization as in the traditional NMF model.

Both of these models are shown to outperform NMF and LDA, which evidences the need for taking advantage of semantic information when considering short texts.

$$\mathrm{C}_{\mathrm{tfidf}} = \mathrm{C}_{\mathrm{tf}} \times \mathrm{idf}(C) \tag{2}$$

$$\mathrm{C}_{\mathrm{tf}} = T \times C \tag{3}$$

$$\mathrm{idf}(C) = \log \frac{M}{\sum_{1 \le d \le M} \mu_{C_t,d}} \tag{4}$$

3.2. 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. (5), 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.

$$\mathrm{PMI}(t) = \frac{2}{T(T-1)} \sum_{i<j \le T} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{5}$$

Topic models that follow probabilistic approaches estimate mixtures over the latent 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. However, perplexity has been shown to not reflect semantic coherence of a topic, sometimes scoring against expert evaluation (Chang et al., 2009).

4. Dataset Definition

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. A total of 94 patients were included in the collection.

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, and self-reported intensity of both pain and disease.

4.1. 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. 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 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.

This preparation pipeline facilitates further processing in two ways, (i) it allows for the study of the patient's integral 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.

4.2. Data challenges

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.

For the first type of challenges, 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. On top of this, 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.

Regarding the second type of challenges, 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, the automatic processing is very limited.

Finally, regarding the third type of challenges, we are concerned with the amount of available data to perform the analysis. If the patient's answers to the 7 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 under the analysis, we would have 94×7 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 notion of a patient could be lost.

5. Experimental Setup

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. This characterization is defined 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 section, 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.

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. Text preprocessing is the task 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. 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. 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 the 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 and is limited by domain adaptability and 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 et al., 2018), which have been pre-trained on Portuguese corpora.

5.1.1. 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 context agnostic or dependent. We evaluate the applied topic models under both of these types of metrics.

For the interpretability metrics, given a fixed number of topics to extract, following the literature, we evaluate the topic coherence of each topic model, as 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. The PPMI metric is defined in Eq. (6), where $t = 10$ is the number of top most weighted words of a topic.

$$\mathrm{PPMI} = \frac{1}{t(t-1)} \sum_{i<j\leq t} \max\{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, 0\} \qquad (6)$$

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, the typical document projection is composed of a highly weighted dimension and the remaining with infinitesimal values. 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, which is not the case. 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. (7), in each topic model, for the number of clusters equal to the number of topics. This metric defines 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.

$$\mathrm{s} = \frac{b-a}{\max(a, b)} \qquad (7)$$

After this evaluation we obtain the most adequate 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 is recovered by aggregating the corresponding projected fragments.

5.2. Characterization

We now define the methodology to visualize and discuss the extracted structures on the latent semantic space, in order to actually compare patients, identify groups, and correlate with demographic and clinical features.

We perform this characterization following three approaches. First, we look at the projected population as a whole, and characterize it. This includes interpretation and labeling of the extracted topics, topic importance mixtures, and the identification of the most common and important topics and words, for the whole population. 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 step allows us to associate types of experiences of pain, according to their descriptions, to specific ranges or values of demographic and clinical parameters. For the third and final step, we split the population into groups defined by the demographic and clinical parameters, and perform the previous analysis in these groups independently. This step allows us to associate values or ranges of demographic and clinical parameters with aspects of experiences of pain.

5.2.1. Topic modeling results and discussion

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 cooccur in the collection of documents. Figure 1 plots the scores for all models, across a wide range of 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. 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.

Regarding the number of topics to extract, 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, 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. Nevertheless, both topic models are still hard to interpret. CluWords (FastText) topics are 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.

These observations allow us to confirm 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. Thus, the topics extracted by CluWords (FastText) represent the most interpretable, well-defined concepts.

Fixing the number of clusters to equal the number of topics, we can observe the concrete silhouettes of each model in Figure 2. The silhouette of the LDA model represents the ideal silhouette of a quality clustering of samples in a given space: due to LDA's statistical inference nature, the lack of instances (documents), and their short length nature, indeed, 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 are put into the same cluster (and, indeed, some of these have scores close to zero), or are poorly assigned to poorly-defined clusters. 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 to factorize the representation matrix, albeit slightly more informative.

According to the previous observations and discussion, we discard the LDA topic space, because the extracted topics are very hard to interpret, their corresponding most weighted words are heavily 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 also 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.

5.2.2. Characterization results and discussion

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 previous section's results and discussion. In this case, it is the one extracted by CluWords (FastText), with $k = 12$ topics, presented before, and repeated here with additional hand labels in Table 1.

There are two important remarks regarding the assigned labels. First, each label is associated with an

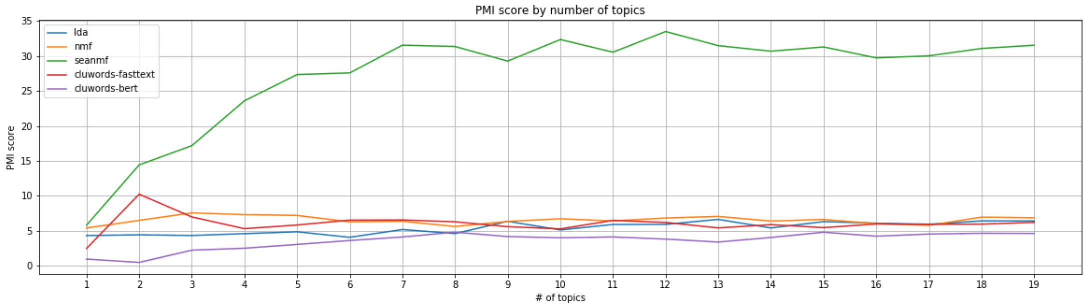

Figure 1: PPMI score of each model, over a range of topics, on the dataset vocabulary.

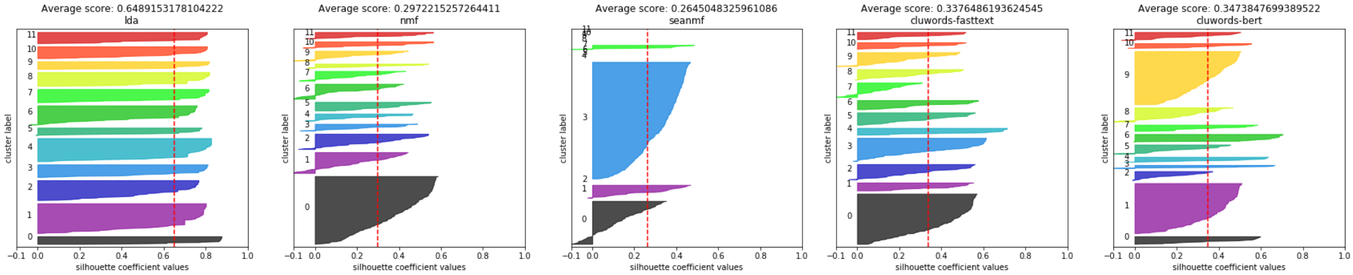

Figure 2: Silhouette of each model, for 12 clusters (equal to the number of topics). Represents the silhouette score assigned to each sample.

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, infeção, 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 1: CluWords(FastText)

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.

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. 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 used interchangeably throughout the whole interview, which makes it hard to reason on their distinction without further exploration.

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 3 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

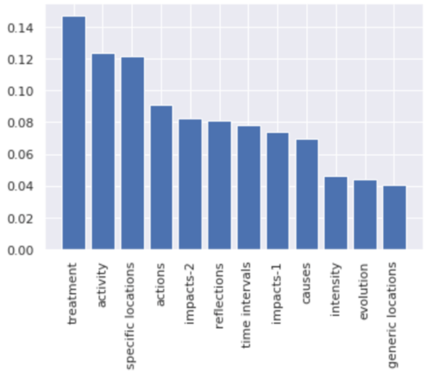

Figure 3: Mean topic mixture of the whole population. Topic weights given as a percentage.

population describing an experience of pain. We can observe a clear elbow for this value of importance, 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 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. This evidence suggests that, for our population, there are aspects of the experience of pain that patients are more commonly inclined to discuss, some even without prompting.

We now define groups of patients that describe their experiences of pain similarly in terms of mixtures of topics. 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. 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 and Wong, 1979). By evaluating the values of inertia, Silhouette Coefficient, Calinski index, and Davies score, across a range of clusters, for this model, because there is no obvious arrangement of patients in well-defined clusters, we decide to cluster the patients in 7 groups. We observe that the cluster's mean mixtures are characterized by a high weight given to one or two topics and small weights scattered across other select few topics. As expected from previous results, the topics that are most commonly assigned high weights, and are used to somewhat easily distinguish each cluster, are the ones that presented more variance and importance, the topics of activity, treatment, and specific locations. Further observations suggest that there is no correlation between the obtained semantic clusters and demographic and clinical parameters.

We now group patients by values or ranges of demographic and clinical parameters. 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 expected, since different pathologies may have different manifestations of pain, including different, more or less specific locations. 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. We do not observe as notable differences with demographic parameter grouping, as with the clinical parameters.

6. Predicting Clinical Parameters

We now 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. This task may be performed on any clinical parameter, however, in this case, we are interested in the diagnosed pathology. This parameter is directly related with the experience of pain, even though the design of the interview, which is the tool used to collected descriptions of pain from patients, was not directly intended for this task. Given the poor distribution across all classes, this experimental setup is only concerned with P1 (41 patients, Rheumatoid Arthritis) and P2 (45 patients, Spondylitis), so that the task is defined as a binary classification task with reasonably balanced classes. Given the limited size of the dataset, it is not be separated into 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 sub-set of $n - 1$ patients and predicting the pathology of the one remaining. All experiments are evaluated by their accuracy on the task.

6.1. 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 summarized in Table 2. The first 4 features are the baselines. The vocabulary-based representations 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*, *full*, and *single question [1-7]*.

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 2: 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.

The *fragment* aggregation looks independently at each of the 7 documents belonging to a patient, as if they were not semantically related.

The *full* 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 vocabulary 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. For these reasons, the results associated with this type of aggregation are expected to be inferior than that of the *fragment* aggregation.

The *single question [1-7]* aggregation presupposes that for the task of pathology classification, 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 pathology, or are simple 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 pathology classification task, all experiments are done in an ablative fashion. This way, each experiment includes all possible permutations of the considered interview questions.

6.2. Results and discussion

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

Table 3: 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 4: Configuration of all experiments.

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 3, resulting in 4 experiments, presented in Table 4. Each experiment encompasses the accuracy of 8 feature types, across 9 types of feature aggregation.

The experimental setup relied on the use of 4 machine learning 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 Support Vector Machine (SVM). For this reason, all results and considerations shown here are in regard to the SVM model with a linear kernel.

The mean accuracy score per experiment configuration allows us to compare experiments in a high-level and to understand the limitations of each aggregation type, in general. The *fragment* type shows higher scores than the *full* 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. 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 this question is informative enough to predict their pathology in our binary classification setting, with a mean accuracy score above 70%. 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 *full* aggregation types.

Focusing on the relevant aggregation types (*fragment*, *full*, *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, 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 syntactic variability of words, may dilute the information carried by the nouns, resulting in a performance score more than 5 percentage points inferior than 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 evident as expected. The removal of stop-words is also reflected in the small difference between EXP. 3 and 4. The following discussion will focus on only these two experiments.

Figure 4 dives 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. Given that the text is already standardized (lemmatization), the TF-IDF features are capable of extracting important information, regardless of of having or not removed stop words, because these are usually assigned very low scores due to their high document frequency nature. Indeed, the superior result obtained with TF-IDF 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 performance of all other models on this task is not evidently different from the NMF baseline. 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.

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 and removed stop words (EXP. 4), considering the *single question (1)* aggregation type. This observation is the main motivation behind the ablative experiment, discussed in the following paragraph.

The 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. We also observe a slight increase in score as we remove more answers that are not from question (1). This is in line with the previous discussion, and can be summarized by the importance of pain location for the diagnosis of these specific pathologies.

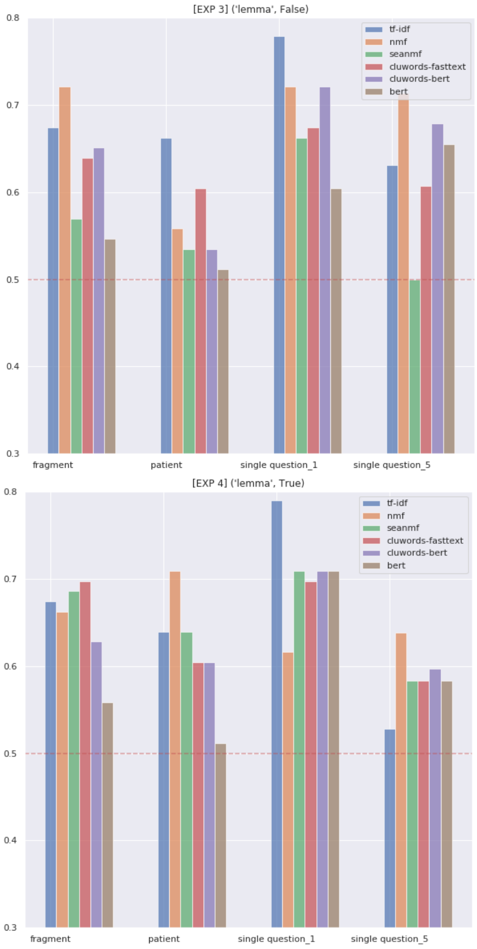

Figure 4: Accuracy score of each feature type, over the different types of feature aggregation. Some baseline results were omitted.

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 Section 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 expression of pain were identified, namely, the emotional state, beliefs, expectations, behavior, and the sociocultural context of the subject.

Based on these observations, in Section 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. 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 Section 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 wasn't 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 collected a parallel dataset of a control group. Indeed, the very definition of a control group, in this context, 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 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.

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

7.2. Linguistic characterization

The linguistic characterization of the population, presented and discussed in Section 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 cooccurrence, 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 by 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 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 clinic parameters. On the other hand, groups of patients given by these external parameters revealed that some groups report slightly different experiences, which suggests to be related to the parameter itself.

7.3. Prediction of clinical parameters

The prediction of clinical parameters presented in Section 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. 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.

7.4. *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.

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 Section 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 extremely 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.

8. Acknowledgments

This work is supported by the 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.

References

Azevedo, L. F., Costa-Pereira, A., Mendonça, L., Dias, C. C., and 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.

Blei, D. M., Ng, A. Y., and 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., Kvarstein, G., and Stubhaug, A. (2008). Assessment of pain. *BJA: British Journal of Anaesthesia*, 101(1):17–24.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 163:1–13.

Dansie, E. and Turk, D. C. (2013). Assessment of patients with chronic pain. *British Journal of Anaesthesia*, 111(1):19–25.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fink, R. (2000). Pain assessment: the cornerstone to optimal pain management. In *Baylor university medical center proceedings*, volume 13, pages 236–239. Taylor & Francis.

Hansen, G. R. and Streltzer, J. (2005). The psychology of pain. *Emergency Medicine Clinics*, 23(2):339–348.

Hartigan, J. A. and 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.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Loeser, J. D. and Melzack, R. (1999). Pain: an overview. *The lancet*, 353(9164):1607–1609.

Melzack, R. (2001). Pain and the neuromatrix in the brain. *Journal of dental education*, 65(12):1378–1382.

Merskey, H. and 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-painorg)*.

Shi, T., Kang, K., Choo, J., and 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*, WWW '18, pages 1105–1114.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 753–761. ACM.

Wilson, D., Williams, M., and Butler, D. (2009). Language and the pain experience. *Physiotherapy Research International*, 14(1):56–65.

Woolf, C. J. (2010). What is this thing called pain? *The Journal of clinical investigation*, 120(11):3742–3744.