APED Associação Portuguesa para o Estudo da Dor

Permanyer

Check for updates

ORIGINAL ARTICLE

Analysis of chronic pain descriptions for base-pathology prediction: the case of rheumatoid arthritis versus spondylitis pathology prediction based on pain descriptions

Análise de descrições de dor crónica para predição de patologia-base: o caso de artrite reumatoide versus espondilite predição de patologia baseada em descrições de dor

Diogo A.P. Nunes[1], Joana Ferreira-Gomes[2], Carlos Vaz[3], Daniela Oliveira[4], Sofia Pimenta[4], Fani Neto[2], and David Martins de Matos[1]*

[1]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal, [2]Department of Biomedicine, Experimental Biology Unit, Faculty of Medicine, University of Porto,Porto, Portugal, [3]Faculty of Medicine, University of Porto, Porto, Portugal, [4]University Hospital Center of São João, Porto, Portugal

Abstract

The language of pain is a sub-language used to describe a subjective, private, and painful experience. In the clinical assessment and management of chronic pain, which is not as straightforward as acute pain, verbal communication is key to convey relevant information to health professionals that would otherwise not be accessible, namely, intrinsic qualities of the painful experience and that of the patient. We raise the hypothesis of applying Natural Language Processing techniques to transcribed verbal descriptions of chronic pain, to capture that information in the form of linguistic features that characterize and quantify the experience of pain of each patient. Furthermore, we demonstrate the application of these features for base-pathology prediction, specifically regarding the diagnosis of rheumatoid arthritis and spondyloarthritis. A dataset of verbal descriptions was collected for this work, considering 85 patients. The descriptions were obtained by having each patient freely answer to an interview of seven questions. The dataset was pre-processed, and features were extracted, which were then fed into binary classification machine learning models. We obtained an accuracy of 79%, in a Leave-One-Out cross-validation fashion. Based on an extensive experimental setup, we conclude that the computational analysis of the language of pain can potentially extract useful information to aid health professionals, in this case, focusing on base-pathology prediction. We also conclude on which semantic features provided more useful information for the task (distribution of pain on the body), and which did not.

Keywords: *Chronic Pain. Computational Pain Assessment. Language of Pain. Natural Language Processing. Prediction.*

Introduction

When there is a manifestation of both disease and pain that is not directly visible to the outside world[1], a verbal interview with a patient is a key moment for pain assessment and management. Here, the information about the cultural, behavioral, and psychosocial dimensions of the subject in pain are conveyed, intentionally or unintentionally, in the form of verbal and non-verbal expressions. The most common expressions of chronic pain are cries, facial expressions, verbal interjections,

***Correspondence:**
Diogo A. P. Nunes
E-mail: diogo.p.nunes@inesc-id.pt

Date of reception: 29-11-2021
Date of acceptance: 02-02-2022
DOI: 10.24875/DOR.M22000014

Available online: 02-03-2022
DOR. 2021;28(2):78-87
www.dor.pt

descriptions, emotional distress, disability, and other behaviors that come as consequences of these. The expression that is the object of study of our work is the spontaneous verbal description of the experience of chronic pain. The verbal description often-times includes valuable information about the bodily distribution of the feeling of pain, temporal patterns of activity, intensity, emotional, and psychological impacts, and others. In addition, the choice of words may reflect the underlying mechanisms of the causal agent[2], which, in turn, can be used to redirect therapeutic processes. Indeed, this forms a specific sub-language which has been studied in the previous research, such as the structuring of the Grammar of Pain[3], and the study of its lexical profile, resulting, namely, in the McGill Pain Questionnaire (MPQ)[4], which is widely used to characterize pain from a verbal standpoint, in clinical settings[5,6]. After one or more interviews, the health professional takes into consideration all of these parameters and interprets the expressions of pain of that patient according to some mental model developed with years of professional and personal experience. All of these studies and tools rely on manual methods and expensive human evaluation.

The computational analysis of syntactic and semantic structures of the language of pain may yield correlations between the content of the descriptions and other relevant medical and non-medical aspects of the painful experience, allowing for a systematic and quantifiable way of characterizing pain and disease manifestations on a linguistic level. This, in turn, has the potential to aid health professionals with the clinical assessment and management of these patients. The automatic qualification and quantification of the language of pain can have multiple applications in the clinical practice, one of them being the prediction of the pathology underlying the experience of pain, which is the focus of this work. This sort of automation can be used as chatbots for patient triage[7], patient-oriented mobile applications with automated statistics from natural language, such as emotion recognition[8], and others. In our case, we envision a system capable of assisting the health professional during the clinical appointment, by providing key statistics automatically extracted from the patient's speech during an interview. Even though this work is focused on diagnostics, various other features and inferences could be made from the description of chronic pain, including pain intensity automatization (absolute, relative, and trend analysis), extraction of pain qualities and quantification of those qualities per patient, patient grouping based on them sharing similar qualities of the experience of pain, and so on.

We raise the hypothesis that Natural Language Processing (NLP) techniques can extract useful information from spontaneous accounts of chronic pain experiences, and that this information, in the form of linguistic features, can be used to aid health professionals. To evaluate this hypothesis, we discuss the process and results associated with its analysis, applying the extracted information to binary pathology prediction, specifically between rheumatoid arthritis and spondylitis. These pathologies were chosen considering the connection with the hospital where the data collection took place, and the balance between the availability of patients and diversity representation for meaningful statistics.

Background

Expression and language of pain

The experience of pain, dependent on its temporal patterns of activity, bodily distribution, and other aspects, is molded by multi-domain cognitive factors, both individual and sociocultural. Some of these, known to influence pain perception and corresponding suffering, are emotional states, beliefs, expectations, and behaviors[9,10]. Language has been found to convey part of this information[2,11]. Understanding how the language of pain is used for expressing specific types of pain experiences allows us to build a linguistic model of pain descriptions. Similar descriptions of pain might describe similar characteristics of different experiences of pain. Characterizing these descriptions by their linguistic features allows us to quantify the relations between different experiences in this abstract, semantic space. Relating these with specific diagnoses allows us to take advantage of the link between linguistic features and clinical pathology.

The work initiated by Melzack and Torgerson[11] aggregated part of the lexical profile of the language of pain, 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 result of these studies was the MPQ, which is widely used to characterize pain from a verbal standpoint, having been demonstrated to provide reliable, valid indices of pain in a relatively efficient way[5]. However, the identified MPQ pain descriptors represent only a portion of the lexical profile of the language of pain. 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 the sociocultural, linguistic background of each patient[6]. Moreover, the MPQ only accounts for the lexical profile of this sub-language, leaving out of consideration more complex structures, such as syntactic and semantic structures (e.g., the Grammar of Pain[3]). In addition, as stated before, all of these studies relied on manual methods and human evaluation. The use of NLP techniques has the potential of overcoming all of these limitations and, thus, extracting more information from descriptions of pain.

Automatic feature extraction

Even though not directly applied to the language of pain, NLP techniques have seen an increase in health-related applications. An important application is the extraction of mental health features from social media texts, showing that informal language can be used to accurately classify social media users as having or not mental health problems[12,13]. Other works focused on extracting entities and relations, such as symptoms, also from social media texts[14,15].

There are various methods to automatically extract linguistic features from text. The extracted features define the feature space, where documents (i.e., strings of text) become embedded. Given a collection of D documents with V unique words (shared vocabulary), traditional methods create feature-vectors based on vocabulary frequencies, and, thus, embed the documents in the vocabulary space. In the Bag-of-Words (BoW)[16] scheme, each document is characterized by a vector of dimension V, where each entry corresponds to the number of times each word of the vocabulary appears in that document. It is called a bag of words because word order is explicitly disregarded. In the Term Frequency/Inverse Document Frequency (TFIDF)[16] scheme, the same bag of words format is used. However, each word frequency count is normalized according to the number of times that word appears in the entire collection. For this reason, TFIDF features are said to be more discriminative: words that are uncommon in the collection of documents have more weight than words that are very common, and, thus, characterize better each document in which they appear.

Other methods, such as topic modeling, focus on extracting latent information in a given document from a collection, embedding it in an abstract topic space. Given a pre-determined number of k topics, these methods estimate the importance of each topic for each document, effectively obtaining a feature-vector of topic distribution for each document. A topic is a distribution of weights over the vocabulary of the document collection, such that the most weighted words are semantically and syntactically related, given the collection. Thus, because of these relations, topics can be interpreted and associated with meaning, usually by giving them representative labels.

Traditional, text-based topic models, such as Latent Dirichlet Allocation (LDA)[17] and Non-Negative Matrix Factorization (NMF)[18], take a vocabulary-based representation of the collection (BoW and TFIDF, respectively), and extract topics following probabilistic and matrix factorization approaches, respectively. However, certain contexts focus on short-text, where the document length shifts from the hundreds of words to the hundreds of characters, such as data from online platforms (e.g., Twitter). In our case, we deal with spontaneous accounts of experiences of pain, which resemble short texts. Extracting topics from short texts presents challenges that the traditional models are not capable of efficiently overcoming. Specifically, word co-occurrence is key information to extract syntactic and semantic relations between words, and topic modeling is, by definition, dependent on the extraction of these relations. However, due to their nature, short-texts have less co-occurrence information, and, thus, pose a greater challenge to traditional topic modeling approaches.

The short-text topic model CluWords[19] exploits external semantic information (i.e., word vectors, also called embedding's, of any given model) by replacing each term in a document BoW representation by a metaword, the CluWord, which represents the cluster of syntactically and semantically similar words, as determined by a pre-trained word-embedding model. This enhanced TFIDF/CluWord representation matrix is then submitted to factorization as in the traditional NMF approach. The main limitation of this model lies in the suitability of external resources to the problem domain.

The short-text topic model SeaNMF[20], on the other hand, overcomes the problems associated with short-text collections by capturing beforehand word and context vectors for the whole vocabulary and explicitly using them instead of the TFIDF representation. This is shown to capture relevant term-context correlation that otherwise would not be fully taken advantage of by the NMF approach. The limitation of this approach lies in the corpus availability and vocabulary diversity.

Materials and methods

Our data are the result of a collaboration project between INESC-ID, Faculty of Medicine of University of Porto, and the Centro Hospitalar Universitário São João, Porto, Portugal (CHUSJ). Dataset collection took place at CHUSJ, from October 2019 to October 2020. The dataset contains spontaneous verbal descriptions of chronic pain experiences, from recorded interviews. The patients are adults (≥ 18-years-old), of either sex, diagnosed with rheumatoid arthritis or spondyloarthritis, and with symptoms of chronic pain. The dataset collection project was approved by the Ethics Committee of CHUSJ. Data confidentiality is explicitly protected: all recordings are anonymous, and results are always presented 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 medically relevant data; such that the patient's personal identification is never used.

Materials: questionnaire

An interview composed of seven questions was designed with the aim of obtaining a natural description of the patient's chronic pain experience, in their own words, while directing it towards the cognitive topics relevant for pain assessment. The script was not pre-validated with patients due to time constraints, but it is the result of several iterations with multiple health professionals that interact with chronic pain patients daily. The script is as follows:

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?

Materials: acquisition setting, challenges, and pre-processing

Interviews with 85 patients were conducted after the pre-scheduled medical consultation with each patient, using a smartphone as a recording device. Due to the setting of the collection, in a clinical context, there are a number of challenges: some patients showed difficulties in verbally expressing themselves; verbal accounts carry speech disfluencies, such as repetition and correction; and the medical office is not free of random environment noises. The data pre-processing step aims at reducing the impacts of these challenges. Interview recordings were submitted to a semi-automatic diarization process, which separates patient audio segments from interview questions. The patient audio segments were manually transcribed, leaving out speech disfluencies. Even though the number of patients is comparable to that of other studies[21,22], it is still an important limitation. To overcome this limitation, we augmented the collection by fragmenting each patient document into the corresponding seven answers to the interview. In this setting, we consider the answers to each question to be semantically independent. Note that patients can still be fully represented by their corresponding seven documents.

We standardize the text through lemmatization (root word extraction) with STRING[23], and remove stop-words (i.e., words that do not provide intrinsic semantic information, such as determinants and pronouns) as defined by the Natural Language Toolkit [24] for Portuguese. Additional words empirically found not to be semantically meaningful for our task, such as the frequent formal address "senhor doutor" (doctor), were also removed. Even though speech disfluencies were explicitly removed, texts resulting from spontaneous speech can still be syntactically incoherent and their linguistic analysis cannot be based on the explicit modeling of syntactic structures. Table 1 shows the ten most frequent words, before and after pre-processing, highlighting the importance of this step.

Table 1. Top 10 most frequent words, before and after preprocessing

Before	*eu* (I), *não* (not), *ser* (be), *que* (which/whose), *ter* (have), *de* (of), *em* (on), *estar* (be), *dor* (pain), *um* (one)
After	*andar* (walk), *dia* (day), *mão* (hand), *poder* (can), *conseguir* (able), *doer* (hurt), *sempre* (always), *medicação* (medication), *joelho* (knee), *pé* (foot)

Table 2. Summary of the four different experiments

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

Table 3. Types of features to extract from the document collection

Feature Type	Dimensions
BoW	$D \times V$
TFIDF	$D \times V$
LDA	$D \times k$
NMF	$D \times k$
SeaNMF	$D \times k$
CluWords (FastText)	$D \times k$
CluWords (BERT)	$D \times k$
BERT (doc2vec)	$D \times k$

D: number of documents in the collection; *V*: size of the vocabulary; *K*: number of extracted topics. BoW: bag-of-words; TFIDF: term frequency/inverse document frequency; LDA: latent dirichlet allocation; NMF: non-negative matrix factorization

Materials: dataset summary

The dataset comprises 85 patients, diagnosed with either rheumatoid arthritis (41) or spondyloarthritis (44). The interview with each patient comprises seven questions. Thus, the dataset considers 85 × 7 = 595 short-text documents. Depending on the context and the needs, a patient can be represented simultaneously by all seven independent documents, by only one of the seven documents, or by all seven documents concatenated together (effectively creating a single, large document).

Methods: type of text

The aim is to extract linguistic features in the form of feature-vectors from the dataset and discover which set of features is most informative for the task of base-pathology prediction, that is, which produces the best accuracy result. This will tell us if verbal descriptions of chronic pain contain valuable information for the assessment and management of chronic pain patients (applied to this task), and how to extract that information automatically.

The first concern is the type of text used as a basis for feature extraction. It can be natural, or lemmatized text, with stop-words removed or not (i.e., applying the pre-processing described in the "Materials: Acquisition setting, challenges, and pre-processing" section, or not). This separates our experimental setup into four experiments, summarized in table 2.

Methods: type of features

Given the setting of each experiment in table 2, we extract a feature-vector for each resulting document (for all *D* documents; table 3). We extract eight types of features: as a baseline, vocabulary-based features (BoW and TFIDF) and traditional topic models (LDA and NMF), and then short-text tuned topic models (SeaNMF, CluWords with FastText embeddings[25], and CluWords with BERT embeddings[26]) and document-embeddings (BERT doc2vec). These are summarized in table 3.

Methods: type of feature aggregation

Given the setting of each experiment in table 2, and one of the extracted feature-vectors according to table 3, each patient is associated to seven independent feature-vectors, corresponding to each of the answers to the interview. To represent each patient with a single vector, the following types of feature-vector aggregation are considered: *fragment, patient, and single question [1-7]* (summarized in table 4).

These aggregations imply different semantics with different feature values.

Fragment looks independently at each of the seven documents belonging to a patient, as if they were not semantically related. In this context, each document is much shorter in length, and semantically focused on less topics (because it has fewer words). By aggregating the seven 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.

Table 4. Types of feature-vector aggregations that represent each patient with a single feature-vector instead of seven vectors

Fragment	The seven vectors are aggregated by their mean value in each dimension.
Patient	The seven documents are concatenated beforehand, resulting in a single document per patient from which the feature-vector is extracted.
Single question [1-7]	Only one of the vectors is considered (corresponding to a single question of the interview).

Patient considers that each patient has a single document (result of concatenating beforehand all seven documents). This means that features are extracted on 85 long documents (equal to the number of patients). However, given that the number of documents is so low, there might be a loss of information, especially regarding word co-occurrence in document windows and complex topic distributions (i.e., a single long document might focus on multiple, disparate topics).

Finally, *single question [1-7]* assumes that for the task of base-pathology prediction, the patient is sufficiently, or better represented by a single question's answer to the entire interview, since there is much less noise. 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. Nonetheless, semantically speaking, the documents are potentially more focused: if there are question's answers in the interview which are prejudicial to the prediction of the base-pathology, or are simply irrelevant, diluting the useful information in noise, this type of aggregation is expected to produce superior results. This aggregation is a simplified version of all the permutations of possible sets of questions to be considered.

Methods: classification models

The experimental setup relies on the use of 4 machine learning models. These are Support Vector Machine (SVM)[27], k-Nearest Neighbors[28], Random Forest[29], and Logistic Regression[30]. The parameters of each model are used as defined by default in the Sci-Kit Learn toolkit[31]. All of these models were used in order to minimize the variability of results, especially those that are intrinsic to the design of the models. After running the experiments for all four models, it was determined that the performance of all models was equal, or inferior, to that of SVM. For this reason, all results and considerations shown after refer to the SVM model.

Methods: evaluation

Given the limited size of the dataset, it is not separated in training and test sets. Instead, evaluation is performed in a Leave-One-Out fashion, for each experiment with each type of feature-vector/aggregation. The score is given as the mean accuracy score of training on every subset of n-1 patients and predicting the base-pathology of the one remaining, until all patients have been selected as test. Experiments are evaluated according to their accuracy (percentage of correct predictions).

Results

Figure 1 shows the scores per experiment (Table 2), across all feature-aggregations (Table 4). These scores are obtained by averaging the accuracy scores of each feature-vector for each experiment. Score variance is also shown. The red dashed line represents the threshold of random binary choice (50%).

This allows us to compare experiments (i.e., type of text) from a high-level and to understand the limitations of each aggregation type, in general. Aggregation types *single question 2, 3, 4, 6, 7*, irrespective of the experiment setting, have accuracy scores below or similar to random choice. For this reason, these are immediately discarded. For the remaining aggregation types (*fragment, patient, single question 1 and 5*), we observe that experiments 1 and 2 have consistently inferior scores. For this reason, these experiments are also discarded.

Figure 2 zooms-in on experiment 3, showing the scores for all baseline feature-vectors, across the selected feature aggregations. Similarly, figure 3 zooms-in on experiment 3, showing the scores for the remaining feature-vectors (short-text models and document embeddings) with the best baselines (TFIDF and NMF) for comparison, across the selected feature aggregations. These figures allow us to compare the importance of each feature-vector type, for each aggregation type, in the setting of experiment 3, that is, with lemmatized text and stop-words not removed.

Figure 4 zooms-in on experiment 4, showing the scores for all baseline feature-vectors, across the selected feature aggregations. Similarly, Figure 5 zooms-in on experiment 4, showing the scores for the remaining feature-vectors (short-text models and document embeddings) with the best baselines (TFIDF and NMF) for comparison, across the selected feature aggregations. These figures allow us to compare the

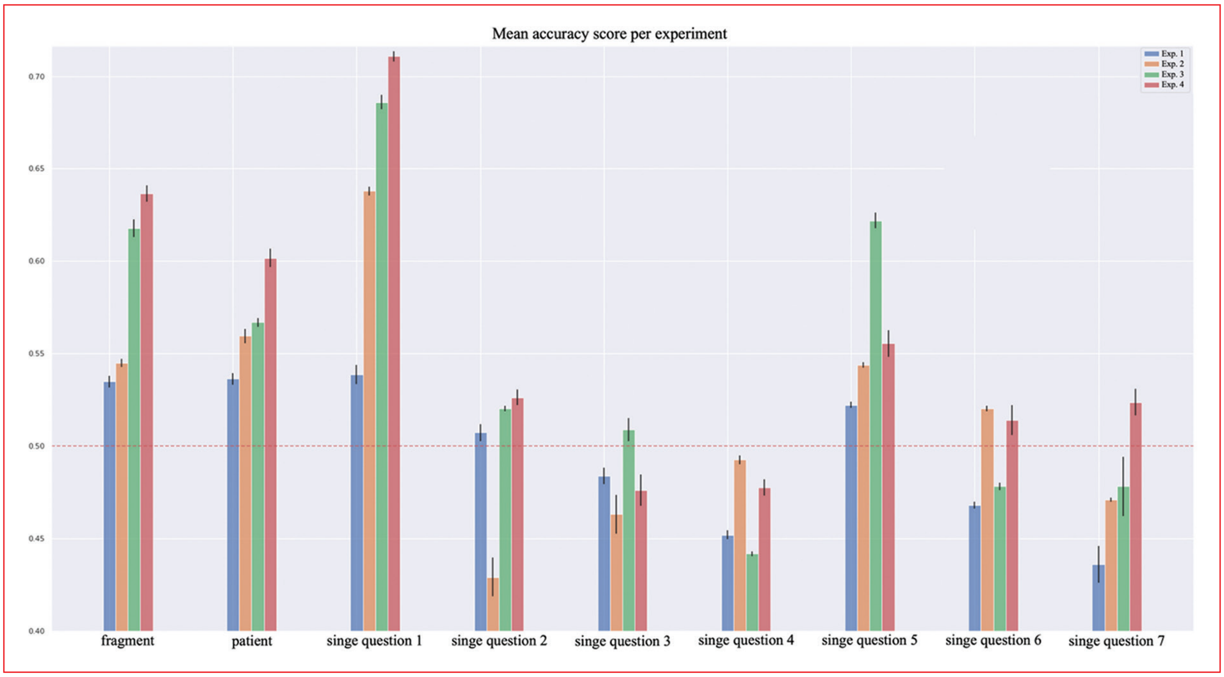

Figure 1. Mean accuracy score of each experiment in Table 2, over the different types of feature aggregation in Table 4.

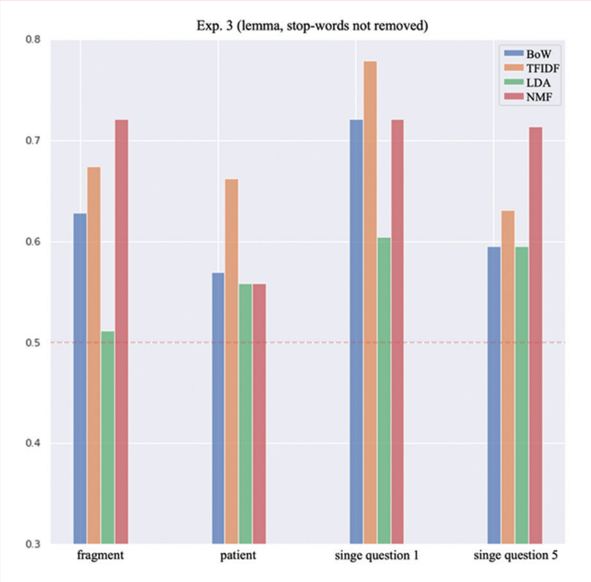

Figure 2. Baseline features (Exp. 3).

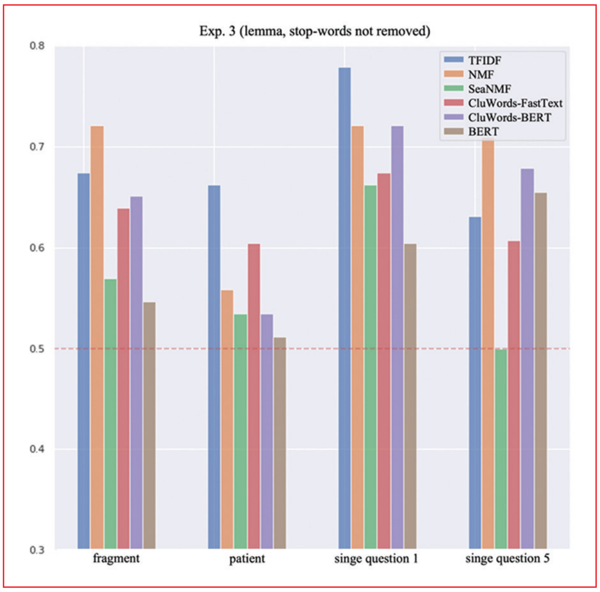

Figure 3. Remaining features (Exp. 3).

importance of each feature-vector type, for each aggregation type, in the setting of experiment 4, that is, with lemmatized text and stop-words removed. Noticeably, in general, classification results are slightly higher for experiment 4.

Discussion

After observing figure 1, we were able to discard aggregation types *single question 2, 3, 4, 6, 7*, because their results, irrespective of the experiment, were below or similar to random choice. Thus, we conclude that,

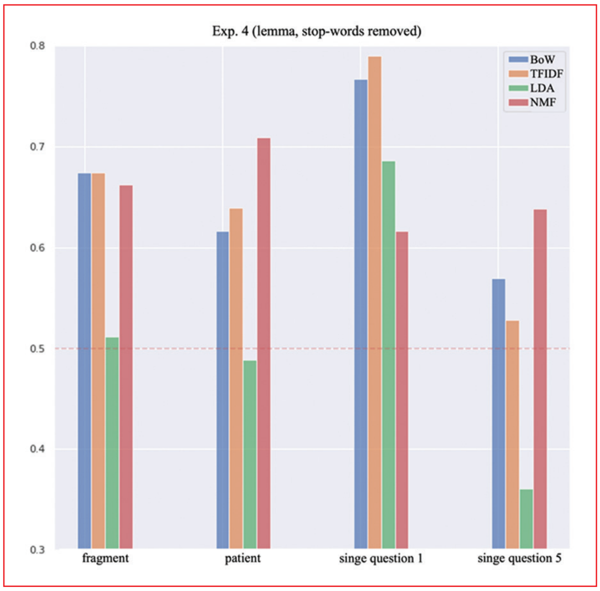

Figure 4. Baseline features (Exp. 4).

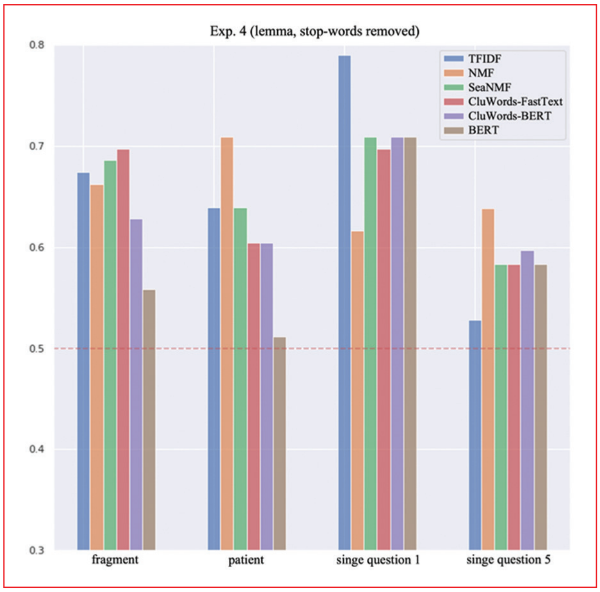

Figure 5. Remaining features (Exp. 4).

by themselves, either the cognitive topics discussed in questions 2, 3, 4, 6, and 7 (pain sensations, patterns of pain intensity, day-to-day and emotional impacts, beliefs regarding treatment quality, and expectation of future pain developments, respectively) of the interview do not convey relevant information for our task, or the patients had considerable difficulty elaborating those topics. Furthermore, for the remaining feature aggregation types, we discarded experiments 1 and 2 because their mean scores are consistently inferior to experiments 3 and 4. With this, we conclude that the setting of experiment 1 results in the poorest performance overall, which can be justified by the fact that it is based on the most raw data (natural text and no stop-words removed), meaning that important information is getting diluted in noise. This is especially evident for *single question 1*, which should be basically a list of nouns (locations of pain 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, is diluting the information carried by those relevant nouns. We also conclude that, overall, using lemmatized text (experiments 3, 4) results in better accuracy scores. The removal of stop-words is also reflected in the small differences between experiments 3 and 4. This discussion validates the proposed pre-processing pipeline in the “Materials: Acquisition setting, challenges, and pre-processing” section. The *fragment* aggregation type displays only slightly higher scores than the *patient* aggregation type, which is unexpected, according to the discussion in the “Methods: Type of Feature Aggregation” section. 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 base-pathology prediction, which is not a trivial task. We can also observe a clear spike in accuracy when using the *single question 1* aggregation type. This means that the patient answers to this question convey relevant information to predict their base-pathology in our binary classification setting, with a mean accuracy score above 70%. This result is in line with clinical literature: even though rheumatoid arthritis and spondyloarthritis share many symptoms (such as joint inflammation, stiffness, and fatigue), an important distinction on the pain location can be made that differentiates the two (rheumatoid arthritis is more incident on joints of the upper limbs, in most cases affecting the wrists, while spondyloarthritis is more incident on back pain and lower limbs)[32-34].

Looking at the vocabulary-based baseline features (BoW and TFIDF), we observe a relevant difference when not removing and removing stop-words (i.e., in figure 2 the scores for these features have a relevant difference, whereas in figure 4 their scores are very similar). Specifically, by removing stop-words, simple word frequency and co-occurrence given by the BoW features becomes as informative as the TFIDF features. As expected, given that the text is standardized (lemmatization), the TFIDF features can extract 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. This is an expected observation due to the limited number of documents and the short-text nature of these documents, which are known to limit the performance of this model. This is also in line with the observations made about the BoW and TFIDF features (LDA is limited by the information carried by the BoW representation, and NMF is limited by the information carried by the TFIDF representation). Both BoW and TFIDF features present higher accuracy scores than NMF, overall (in some cases, with an increase of almost 20% points). However, it is important to note that when referring to the *fragment* aggregation type, there is no evident distinction between these three types of features. Indeed, this suggests that for the task of binary base-pathology prediction, a listing of pain locations (*single question 1*) is more informative than any other type of elaboration on the patient's pain manifestation. The same reasoning applies to all other topic models, which scores are plotted against the best baselines in Figures 3 and 5. 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 suitability of the pre-trained model to the context of our data, and the infeasibility of model adaptation due to the shortage of data.

Finally, we observe that for our setting of binary base-pathology prediction (specifically, between rheumatoid arthritis and spondyloarthritis), the TFIDF features are, overall, the best information extraction method, with an absolute score of 79% with lemmatized text and removed stop-words (the setting of experiment 4), given the *single question 1* aggregation type (Fig. 5).

Conclusion and future work

In this work, we have analyzed the hypothesis that the language of chronic pain conveys relevant information for the assessment and management of these patients, and that NLP techniques can extract this information. Specifically, for base-pathology prediction, we analyzed and discussed which linguistic features extract the most relevant information and determined which parts and aggregations of chronic pain descriptions actually conveyed relevant information for accurate base-pathology prediction.

We conclude that verbal descriptions of chronic pain experiences convey relevant information for the assessment and management of these patients, specifically in the case of base-pathology prediction. We also conclude that NLP techniques can successfully extract this information. Based on an extensive experimental setup, we identified the most relevant carrier of information for our task: having patients describe the location distribution of their pain (the first question of the interview). Moreover, we conclude that the TFIDF features are, overall, the best method to extract that information, with an accuracy of 79% with lemmatized text and removed stop-words.

Future work, on a first stage, should focus on expanding the dataset with more patients. This will allow for evaluation methods other than Leave-One-Out and offer better generalizations, specifically by having a set of patients for model training and a distinct, large set of patients for model evaluation. This would allow for the evaluation of prediction accuracy, precision, recall, and others[35]. Additionally, it will enrich the vocabulary of the dataset and possibly offer better results. Finally, other future considerations should include descriptions of chronic pain obtained without an interview, because this will intrinsically reveal the relative importance of the cognitive topics for each patient whilst removing any possible biases from the prompting.

Acknowledgments

Diogo A.P. Nunes is supported by a scholarship granted by Fundação para a Ciência e Tecnologia (FCT), with reference 2021.06759.BD. Additionally, this work was supported by Portuguese national funds through FCT, with reference UIDB/50021/2020.

References

1. Loeser JD, Melzack R. Pain: an overview. Lancet. 1999;353:1607-9.
2. Wilson D, Williams M, Butler D. Language and the pain experience. Physiother Res. Intl 2009;14:56-65.
3. Halliday MA. On the grammar of pain. Funcs Lang. 1998;5:1-32.
4. Melzack R. The McGill pain questionnaire: major properties and scoring methods. Pain. 1975;1:277-99.
5. Katz J, Melzack R. Measurement of pain. Surg Clin North Am. 1999;79:231-52.
6. Sullivan MD. Pain in language: from sentience to sapience. Pain Forum. 1995;4:3-14.

7. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Inform Med. 2020;28(4):241-7.
8. Koolagudi SG, Rao KS. Emotion recognition from speech: a review. Int J Speech Technol. 2012;15:99-117.
9. Hansen GR, Streltzer J. The psychology of pain. Emerg Med Clin North Am. 2005;23:339-48.
10. Azevedo LF, Costa-Pereira A, Dias LM, Castro-Lopes JM. Epidemiology of chronic pain: a population-based nationwide study on its prevalence, characteristics and associated disability in Portugal. J Pain. 2012;13:773-83.
11. Melzack R, Torgerson W. On the language of pain. Anesthesiology. 1971;34:50-9.
12. Yates A, Cohan A, Goharian N. Depression and Self-harm Risk Assessment in Online Forums. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017. p. 2968-78.
13. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. Proceedings 27th International Conference on CL (COLING 2018). ACL; 2018.
14. Foufi V, Timakum T, Gaudet-Blavignac C, Lovis C, Song M. Mining of textual health information from reddit: analysis of chronic diseases with extracted entities and their relations. J Med Internet Res. 2019;21: e12876.
15. Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T. What patients can tell us: topic analysis for social media on breast cancer. JMIR Med Inform. 2017;5:e23.
16. Manning CD, Schutze H. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: MIT Press; 1999.
17. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993-1022.
18. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788.
19. Viegas F, Canuto S, Gomes C, Luiz W, Rosa T, Ribas S, et al. Cluwords: exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. Proceedings 12th ACM International Conference WSDM'19; 2019. p. 753-61.
20. Shi T, Kang K, Choo J, Reddy CK. Short-text Topic Modeling Via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. Proceeding WWW'18; 2018. p. 1105-14.
21. Carlson LA, Jeffery MM, Fu S, He H, McCoy RG, Wang Y, et al. Characterizing chronic pain episodes in clinical text at two health care systems: comprehensive annotation and corpus analysis. JMIR Med Inform. 2020;8:e18659.
22. Lascaratou C. The Language of Pain: expression or Description? The Netherlands: John Benj Publishing Company; 2007.
23. Mamede NJ, Baptista J, Diniz C, Cabarrão V. STRING: an Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. PROPOR 10th International Conference on Computational Processing of Portuguese; 2012.
24. Bird S, Loper E, Klein E. Natural Language Processing with Python. Sebastopol, California: O'Reilly Media Inc.; 2009.
25. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. ACL. 2017;5:135-46.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proc ACL. 2019;1:4171-86.
27. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273-97.
28. Mucherino A, Papajorgji PJ, Pardalos PM. k-Nearest Neighbor Classification. New York: Springer; 2009.
29. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
30. Wright RE. Logistic regression. In: Grimm LG, Yarnold PR, editors. Reading and Understanding Multivariate Statistics. Washington, DC: APA;1995. p. 217-44. 1995. p. 217-44.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825-30.
32. Rindfleisch AJ, Muller D. Diagnosis and management of rheumatoid arthritis. Am Fam Phys. 2005;72:1037-47.
33. Rojas-Vargas M, Muñoz-Gomariz E, Escudero A, Font P, Zarco P, Almodovar R, et al. First signs and symptoms of spondyloarthritis data from an inception cohort with a disease course of two years or less (REGISPONSER-Early). Rheuma. 2009;48:404-9.
34. Mease PJ, Liu M, Rebello S, Kang H, Yi E, Park Y, et al. Comparative disease burden in patients with rheumatoid arthritis, psoriatic arthritis, or axial spondyloarthritis: data from two corrona registries. Rheum Ther. 2019;6:529-42.
35. Russell SJ. Artificial Intelligence: a Modern Approach. Upper Saddle River, New Jersey: Pearson Education, Inc.; 2010.