Intensity Prediction over Health-related Quality-of-Life Variables Extracted from Self-reported Patient Narratives

Tanjeb Tawhid[1], Philipp Cimiano[1,2], and Matthias Hartung[1]

[1]Semalytix GmbH, Bielefeld, Germany
[2]Bielefeld University, Faculty of Technology, Germany

Abstract

Health-related Quality of Life (HRQoL) plays a pivotal role in patient-reported outcomes, and thus in regulatory drug approval and health technology assessment contexts. Self-reported patient narratives from social media provide a rich source of information in order to extract HRQoL-related information which may complement established research instruments or even overcome some of their shortcomings. In this paper, we present work on automatically assigning intensity scores to HRQoL variables extracted from such narratives. Intensity captures the subjectively perceived importance of a particular HRQoL dimension for a patient's well-being, thus fostering a better understanding about how individuals are impaired by a disease or disability in their daily life. We present two approaches towards intensity prediction based on text classification and ranking algorithms. Our experiments show that both approaches provide viable solutions, with F_1 scores of up to 0.80 under a pairwise formulation of the problem. In a comparative end-to-end evaluation on a discrete output scheme, both approaches are on par, with a potential advantage for the ranking model due to a reduction in annotation complexity.

Introduction

Within the emerging paradigm of patient-focused drug development[1], patient-reported outcomes (PRO) are gaining increasing relevance in regulatory drug approval and healthcare in order to demonstrate particular benefits of health interventions from the patients' perspective [1]. Health-related quality-of-life (HRQoL), defined by the WHO as comprising physical, mental, and social well-being [2], broadens the perspective of health beyond the confined settings of clinical trials, and therefore plays a pivotal role in PRO measurement [3].

As a latent concept, HRQoL cannot be directly observed, but needs to be operationalized via standardized multidimensional instruments such as surveys or structured interviews which are designed as to enable long-term follow-up by monitoring changes in self-reported HRQoL during the patient journey [4]. Inherent limitations of such instruments may concern blind spots due to (i) survey questions being limited to pre-defined items, and (ii) differences in the subjectively

[1]`https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical`

Table 1: HRQoL domains and facets incorporated within each domain according to [2]

Domain	Facets
Physical Health	Energy & Fatigue; Pain & Discomfort; Sleep & Rest
Psychological Health	Body Image and Appearance; Positive Feelings; Negative Feelings; Self-esteem; Thinking, Learning, Memory & Concentration
Level of Independence	Mobility; Activities of Daily Living; Dependence on Medicine; Work Capacity
Social Relations	Personal Relationships; Social Support; Sexual Activity
Environment	Financial Resources; Freedom, Physical Safety & Security; Health & Social Care; Home Environment; Opportunities for Acquiring New Information and Skills; Recreation & Leisure; Physical Environment; Transport

perceived relative importance of individual items or entire dimensions of HRQoL, and longitudinal changes of these differences [5]. In the worst case, the above issues can mutually reinforce each other such that continuously asking respondents the same questions may lead to getting the same answers without substantial information increase, and possibly even without noticing that the intensity of a seemingly unchanged HRQoL impairment increases over time, because the importance of the respective facet has increased, e.g., as a consequence of disease progression or changes in the treatment regime [6].

Our work addresses both these challenges by algorithmically analyzing patient-reported narratives from social media sources based on natural language processing and text analytics. Our goal is to detect mentions of HRQoL concepts (following the WHO taxonomy as sketched in Table 1) in such narratives and link them to self-reported disease burdens and treatment outcomes from the same sources. Thus, we are working towards a scalable qualitative research instrument for monitoring real-world patient needs through the lens of HRQoL concepts extracted from unsolicited narratives on social media.

In this paper, we focus on the task of augmenting HRQoL concepts with indicators of their subjectively perceived intensity, where intensity captures the subjectively perceived importance that a patient assigns to a particular HRQoL facet. Depending on application and context, intensity can be modeled as real-valued scores or discrete labels. In this work, we adopt the latter option: Each HRQoL facet is assigned one of the three intensities *low/basic/high importance*, signifying the magnitude of patients being affected in a particular HRQoL facet.

As our main contribution, we compare two supervised machine learning approaches towards intensity prediction, i.e., text classification and ranking. While text classification can be considered as the obvious and straightforward approach to the problem, annotating text for discrete intensities may be quite challenging and cognitively demanding. A more lenient approach from the annotation perspective is to frame the task as a ranking problem. Under this formulation, given pairwise preference annotation, a machine learning model can be trained to predict real-valued scores for given texts, which will subsequently be decoded into discrete categories. Based on the hypothesis that the pairwise nature of the ranking problem (compared to the three-class schema in discrete modeling) may reduce the complexity of the task, we are interested in comparing the two approaches with regard to their end-to-end predictive performance.

Related Work

Intensity prediction has gained growing attention in the NLP community, largely due to SemEval-2018 Task 1: Affect in Tweets [7]. The task consists of five subtasks where machine learning systems need to infer the intensity of emotion or sentiment felt by a person from their tweets in terms of different output schemes (real-valued scores, ordinal classes, or discrete categories). Contrary to our work, the SemEval tasks consider different target variables (emotions or sentiment, vs. HRQoL facets), and all annotations provided are in line with the respective task's output scheme.

The best-performing models in the competition, SeerNet [8] and PlusEmo2Vec [9], follow similar approaches based on various predictive models that are trained on independent feature sets (some of them derived from pre-trained neural networks, others from task-specific lexical resources) and subsequently combined into ensemble models. Since the classifications datasets are identical to the regression ones, PlusEmo2Vec does not use separate classification models but utilizes thresholds computed from the training data or learned through polynomial regression to map the regression scores to ordinal classes. Our work adopts a similar idea for mapping ranking scores to discrete intensity classes, based on a fully unsupervised approach, though.

Zhang et al. [10] use the pre-trained BERT model with a task-specific output layer, but without incorporating any specialized embeddings or lexical features. Despite its simplicity, the model achieves good performance, even outperforms PlusEmo2Vec on two subtasks. Likewise, our work also capitalizes on transfer learning, using the pre-trained RoBERTa model [11] as a basis.

Methods for Intensity Prediction

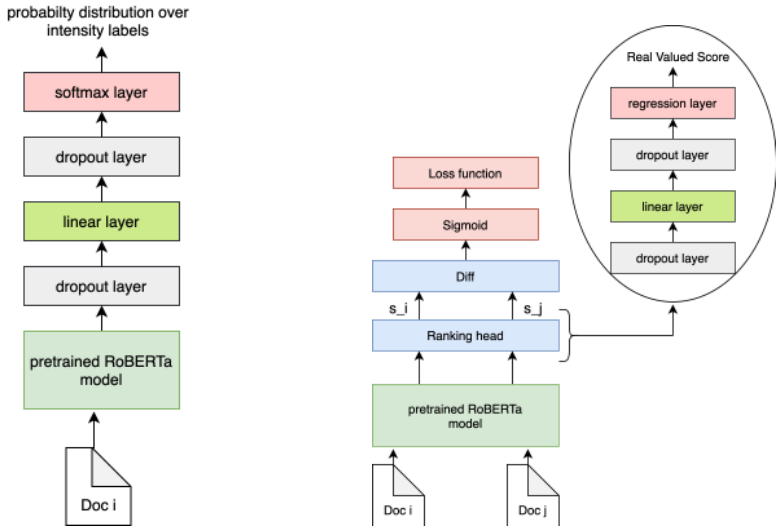
We compare two formulations of the intensity prediction task, either as a text classification or a ranking problem. In the following, both approaches are described in terms of model architecture, training and inference procedures. A high-level summary of both approaches is given in Fig. 1.

Text Classification for Intensity Prediction

Though a variety of models can be used for classification, such as decision trees, support vector machines, and neural networks, following the success of pretrained transformer language models, we use the RoBERTa model [11] for our purpose. Since RoBERTa is trained with a masked language modeling objective, to facilitate classification, we add a classification head which consists of a linear layer with *tanh* activation, squished between two dropout layers and a softmax projection layer on top (cf. Figure 1a). For training, we rely on cross-entropy loss which is computed based on the output of the softmax layer.

Pairwise Ranking for Intensity Prediction

Our ranking-based approach consists of (i) a ranking module which implements a combination of RankNet [12] with a pre-trained RoBERTa model (denoted as Rankformer) to predict ranking scores, and (ii) a decoder which categorizes these scores into discrete intensity labels.



(a) Classification model (b) Ranking model

Figure 1: Model architectures: classification and ranking

Rankformer. Figure 1b depicts the Rankformer model, consisting of a ranking head on top of the pretrained RoBERTa language model. The ranking head is identical to the classification head except for having a regression layer on top instead of a softmax layer. The Rankformer takes a pair of documents and their pairwise preference as input. Both documents are passed through the pretrained RoBERTa model to obtain respective representations. Following [11], we take the output at the first token as our document representation. The document vectors are then passed to the ranking head to compute respective ranking scores s_i and s_j. Regarding the cost function, following RankNet [13], we take the sigmoid over the difference between the two scores, followed by computing the cross-entropy loss with respect to the true pairwise preference. As an alternative loss function, we also experiment with the pairwise hinge loss [14].

Decoder. The ranking model only predicts a real-valued score for any given document; we still need to map the score to one of the intensity levels. In the absence of labeled data, we frame the problem as an unsupervised clustering task using Gaussian Mixture Models (GMMs). A GMM has two types of parameters: the mixture components' weights (π_k) and the components' means (μ_k) with their covariances (Σ_k). Following [15], we estimate these parameters using Expectation Maximization. Based on the hypothesis that higher ranking scores should signal higher intensity, we map each GMM component k to an intensity level based on its component mean μ_k in decreasing order.

Table 2: Discrete annotation examples with intensity levels high/basic/low importance

High	Now, hopefully this drug will not give me any side effects like the awful joint pain that the Remicade was giving me and I can be ALMOST normal.
Basic	I had a week without pain or being sick to my stomach.
Low	It wasn't as bad as I was expecting, hardly hurt at all and I'm proud that I could inject myself.

Experiments and Results

Experimental Settings and Data Sets

HRQoL incorporates different facets of an individual's quality of life (cf. Table 1). For the scope of this study, we focus on the facets PAIN & DISCOMFORT, MOBILITY, NEGATIVE FEELINGS, POSITIVE FEELINGS, and FINANCIAL RESOURCES. For each facet, we annotate data both in a discrete and pairwise manner. Data points for annotation are sampled from a corpus comprising anonymized patient narratives in the English language from health-related, publicly accessible social media forums and blogs. Upstream machine learning models and knowledge graph tagging are used to identify content authored by patients, as well as sentences referring to one of the HRQoL facets mentioned above. Reject options are included in each annotation task to address cases of HRQoL facets being erroneously selected by an upstream model. Annotation is done at the sentence level from the perspective of self-reported importance; i.e., we evaluate the subjectively expressed importance of the respective variable to a patient's life.

Discrete Annotation. For the discrete case, each text is annotated for one of the intensity levels *low/basic/high importance*. Assuming that patients' narratives in social media are manifestations of the communicative principle of relevance [16], we assign *basic importance* to texts which contain a mention of the corresponding variable. For high importance, we look for linguistic cues emphasizing the variable. Similarly, for low importance, we look for cues that downplay the effect of the variable. Table 2 provides examples for each class.

Pairwise Annotation. In the pairwise task, annotators are instructed, given a pair of patient-authored narratives, to judge if one expresses higher importance with respect to a certain predetermined QoL facet than the other. Concretely, given a text pair (A, B) and a facet Q, the label 1 is assigned to the pair if the patient-reported importance of Q in A is higher than in B, or 0 otherwise. Table 3 provides examples for each case.

Annotated Data Sets. Table 4 presents an overview of the annotated data sets resulting from discrete and pairwise annotation. These annotations are used as follows: The pairwise annotations are used for training and validating the ranking model. The discrete annotations are separated for training and testing according to an 80%/20% split. The test split is used for evaluating both model types as discussed in the experiments below.

Table 3: Pairwise annotation examples with pairwise intensity labels

1	text A	I cant take it because of my nose bleeds.
	text B	I was only on pain meds for about 3 1/2 days following the surgery and I haven't needed them since.
0	text A	I have mainly sustagen at the moment (I think my crohns is flaring because I don't have much of an appetite, which usually happens in a flare), and with the sustagen, I add sugar.
	text B	Also, ne of my fallopian tubes was completely adhered to my small bowel (explains some of the "female" pain)

Table 4: Number of annotated data points per data set and HRQoL facet

	Pain & Discomfort	Mobility	Financial Resources	Negative Feelings	Positive Feelings	IAA (Cohen's κ)
Discrete	1412	223	576	618	3280	0.317
Pairwise	1267	324	352	885	1134	0.324

Training and Evaluation

We train the classification model on the discrete data and the ranking model on the pairwise data using Adam[2] for optimization. Since both models are quite similar from an architectural point of view, we use the same set of parameters[3] except for the number of epochs trained which varies with data set size. Following the sequential transfer learning paradigm, we take a fine-tuning approach to train the models by adapting the pre-trained representations to the task-specific data. Regarding mixture model training, we use the default settings provided by scikit-learn[4] except for the number of mixture components, which is set to the number of intensity categories. For evaluation, we use macro-averaged F_1 score per output class. In our experiments, we found the ranking models trained with pairwise hinge loss to outperform their counterparts trained on cross-entropy loss. Therefore, all ranking results reported below are based on the former.

Experiments and Results

Discrete Classification. Motivated by our application context, we are particularly interested in ways to generate discrete output labels for the intensity task. To this end, both models are compared on the test set annotated with discrete intensity scores. In order to evaluate the ranking model on discrete data, single documents are fed to the ranking model and the ranking scores are categorized into discrete classes. The results for this experiment are presented in Table 5 (second and third column). Despite a natural concurrence of the task with the classification model, both models turn out to be largely on par at an averaged F_1 score of 0.61.

[2] `https://keras.io/api/optimizers/adam/`

[3] Classification/ranking head input size: 768; Classification/ranking head linear layer size: 3072; Dropout: 0.1; Batch size: 16; Adam ϵ: 1e-8; Adam β_1: 0.9; Adam β_2: 0.98; Learning rate: 2e-5; Learning rate decay: linear; Warmup ratio: 0.06.

[4] `https://scikit-learn.org/stable/modules/mixture.html`

Table 5: Classification and Ranking results (macro-averaged F_1 scores per HRQoL facet)

HRQoL Facet	Discrete		Pairwise	
	Clf.	Ranking	Clf.	Ranking
Pain & Discomfort	0.62	0.48	**0.71**	**0.71**
Mobility	0.55	0.74	0.59	**0.82**
Financial Resources	0.85	0.77	**0.93**	0.82
Negative Feelings	0.41	0.54	0.59	**0.84**
Positive Feelings	0.61	0.50	**0.84**	0.79
Average	0.61	0.61	0.73	**0.80**

Pairwise Ranking. For pairwise evaluation of both models, we transform the discrete test data into pairwise data by generating all possible pairs with different labels. As can be seen from Table 5 (fourth and fifth column), the ranking model yields a macro-averaged performance of F_1=0.80. With regard to individual facets, ranking performance is relatively constant, ranging from F_1=0.84 for NEGATIVE FEELINGS to F_1=0.71 for PAIN & DISCOMFORT. Since the classification models expect single text as input, pairwise predictions are derived from the discrete output by comparing individual predictions for each pair of the transformed pairwise data. This leads to an average performance of F_1=0.73 across facets (with FINANCIAL RESOURCES as a remarkable outlier), which indicates that, even though the classification model is tailored towards detecting finer-grained differences, it lags behind on the ranking task.

Discussion. The experiments reported here suggest that algorithmic intensity prediction on QoL facets does not necessarily lean itself towards discrete modeling, despite its proximity to ordinal scales as they are often used in established research instruments. In fact, our results suggest to reduce the task to a pairwise ranking problem. However, the fully unsupervised decoding approach used in these experiments prevent the ranking model from unfolding its full potential on the discrete task. Future work should explore ways to jointly optimize the ranking and the mixture model. As a potential advantage, the pairwise approach reduces the complexity of the underlying annotation task to binary decisions; however, comparing inter-annotator agreement scores for both annotation tasks reflect this effect only marginally (cf. Table 4). Interestingly though, the pairwise ranking model shows a negative correlation with the number of available pairwise annotations per HRQoL facet (Pearson's ρ=-0.67), whereas such a correlation is not observable for the discrete classifier and discretely labeled data points (ρ=0.04). While this effect clearly needs closer investigation, it might be a hint into the direction that the ranking model, besides reducing the task *complexity*, also requires smaller *volumes* of annotated data.

Conclusion

In this paper, we have introduced the new task of predicting intensity scores for HRQoL variables extracted from patient-reported narratives in social media, along with a comparative evaluation of NLP model architectures towards first effective solutions.

While HRQoL intensity may facilitate deeper understanding about how an individual's well-being is affected by a disease or disability as well as evaluation of corresponding treatment outcomes and identification of unmet medical needs, it is usually not directly observable from established survey-based HRQoL instruments. Thus, our study supports the argument that social media data can be a valuable source to extract unsolicited patient-reported outcomes from large online populations [17, 18].

Acknowledgement

This work was partially funded by the German Ministry for Education and Research (BMBF) under grant no. 01IS19080C.

References

[1] Cappelleri JC, Bushmakin AG, Alvir JMJ. Patient-Reported Outcomes: Development and Validation. In: Alemayehu D, Cappelleri JC, Emir B, Zou KH, editors. Statistical Topics in Health Economics and Outcomes Research. CRC Press; 2018. p. 15–46.

[2] The WHOQOL Group. The World Health Organization Quality of Life assessment: position paper from the World Health Organization. Soc Sci Med. 1995 Nov;41:1403–1409.

[3] Bullinger M, Quitmann J. Quality of life as patient-reported outcomes: principles of assessment. Dialogues Clin Neurosci. 2014;16:137–145.

[4] Daig I, Lehmann A. Verfahren zur Messung der Lebensqualität [Procedures to Measure Quality of Life]. Z Med Psychol. 2007;16:5—23.

[5] Bernhard J, Lowy A, Mathys N, Herrmann R, Hürny C. Health related quality of life: A changing construct? Qual Life Res. 2004;13:1187—1197.

[6] Leuteritz K, Richter D, Mehnert-Theuerkauf A, Stolzenburg JU, Hinz A. Quality of Life in Urologic Cancer Patients: Importance and Satisfaction with Specific Quality of Life Domains. Preprint under review. 2021.

[7] Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S. Semeval-2018 task 1: Affect in tweets. In: Proc. of the 12th Int. Workshop on Semantic Evaluation; 2018. p. 1–17.

[8] Duppada V, Jain R, Hiray S. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. arXiv preprint arXiv:180406137. 2018.

[9] Park JH, Xu P, Fung P. Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags. arXiv preprint arXiv:180408280. 2018.

[10] Zhang L, Huang HL, Yu Y, Moldovan D. Affect inTweets: A Transfer Learning Approach. In: Proc. of the 12th Language Resources and Evaluation Conference; 2020. p. 1511–1516.

[11] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.

[12] Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning to rank using gradient descent. In: Proceedings of ICML; 2005. p. 89–96.

[13] Burges CJ. From RankNet to LambdaRank to LambdaMART: An overview. Learning. 2010;11(23-581):81.

[14] Chen W, Liu TY, Lan Y, Ma Z, Li H. Ranking Measures and Loss Functions in Learning to Rank. In: Proceedings of NIPS; 2009. .

[15] Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.

[16] Sperber D, Wilson D. Relevance: Communication and Cognition. Blackwell; 1986.

[17] Cook NS, Kostikas K, Gruenberger JB, Shah B, Pathak P, Kaur VP, et al. Patients' Perspectives on COPD: Findings from a Social Media Listening Study. ERJ Open Res. 2019.

[18] Patalano F, Gutzwiller FS, Shah B, Kumari C, Cook NS. Gathering Structured Patient Insight to Drive the PRO Strategy in COPD: Patient-Centric Drug Development from Theory to Practice. Adv Ther. 2020;37:17–26.