From Pain to Praise: Aspect-Based Sentiment Analysis for Norwegian Patient Feedback

Lilja Charlotte Storset[1], **Elma Jelin**[2], **Rebecka Maria Norman**[2], **Øyvind Andresen Bjertnæs**[2], **Lilja Øvrelid**[1], **Erik Velldal**[1]
[1]University of Oslo, [2]Norwegian Institute of Public Health
{liljacs,liljao,erikve}@uio.no
{elma.jelin,rebecka.norman,oyvindandresen.bjertnaes}@fhi.no

Abstract

This paper describes a new dataset for aspect-based sentiment analysis (ABSA) for analyzing patient feedback about healthcare services. In an interdisciplinary collaboration spanning the fields of natural language processing and healthcare research, we manually annotate a dataset of 2382 free-text comments collected from national patient experience surveys in Norway, covering two sub-fields of services – special mental healthcare and general practitioners. Annotations are provided on both the sentence- and comment-level, covering a fine-grained set of 25 unique healthcare-related aspects and their polarities. We also report results for fine-tuning both encoder- and decoder models on the resulting dataset, comparing different modeling strategies, like joint and sequential prediction of aspects and polarity. The resources developed in this work can assist healthcare researchers in the analysis of patient feedback, bringing a much more efficient approach compared to today's manual analysis, potentially leading to improved patient satisfaction and clinical outcomes.

1 Introduction

In most countries, patients can express their experiences with healthcare services through platforms such as online review websites, patient forums and surveys. These sources provide valuable insights into a range of factors influencing patient satisfaction. However, manual analysis of large volumes of such data is both time-consuming and resource-intensive, making it difficult to capture trends regarding patient concerns in a timely manner.

While natural language processing (NLP) has been widely used for automating analysis of large volumes of feedback in other fields, such as e-commerce and hospitality (Nazir et al., 2022), automated analysis has not been as widely adopted the within the field of healthcare as of yet. Recently, however, there has been a rise of research on incorporating NLP tools in analysis of patient feedback (Jelin et al., 2025). While earlier healthcare studies using sentiment analysis largely have focused on overall polarity, aspect-based sentiment analysis (ABSA) offers a more fine-grained alternative, potentially giving more nuanced and precise insights into patient narratives.

A successful adoption of ABSA in the healthcare domain requires an interdisciplinary approach, bridging advances in NLP with expertise from healthcare research. This paper describes an extension to the Norwegian Patient Comment corpus introduced by Mæhlum et al.; Rønningstad et al. where we add annotations for aspect-based sentiment analysis; $\text{NorPaC}_{\text{absa}}$. The data is based on free-text comments collected from surveys conducted by the Norwegian Institute of Public Health (NIPH) to form part of their so-called patient-reported experience measures (PREMs). More specifically, our data focuses on feedback regarding special mental healthcare and general practitioners, where healthcare researchers have defined and annotated a rich set of aspects and their corresponding positive/negative polarity. These cover a range of concerns related to topics including healthcare providers, organization of health services, availability, environment and facilities, and more. The paper also reports on a series of modeling experiments on the data, comparing different model architectures and data granularities.

2 Background

In this section we first give a brief discussion of previous work on aspect-based sentiment analysis, before providing some more background on the national patient experience surveys conducted by the Norwegian Institute of Public Health (NIPH), which forms the starting point of the current work.

2.1 Related Work on ABSA

Aspect-based sentiment analysis (ABSA) as a formalized subfield of sentiment analysis (SA), gained considerable attention from the NLP research community with the SemEval shared tasks (Pontiki et al., 2014, 2015, 2016). While common methods in this phase mostly included more traditional supervised machine learning with manual feature engineering, typically implemented as pipeline architectures (Pontiki et al., 2014), the most widely used methods for ABSA have since shifted towards neural approaches. Particularly common are transformer-based encoder models such as BERT (Vaswani et al., 2017). These models are often fine-tuned end-to-end for ABSA, or used in pipelines that separate aspect- and polarity classification. Some ap-

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 191–202

proaches adopt hybrid architectures, combining BERT-based representations with models such as RNNs and CNNs (Chauhan et al., 2023). Further, generative decoder models such as GPT (Brown et al., 2020) have become increasingly popular for end-to-end ABSA, often applied in zero-shot or few-shot settings, but occasionally also through fine-tuning (Simmering and Huoviala, 2023).

Compared to consumer- and product review domains, there has not been much research on ABSA within healthcare, and none specifically for Norwegian healthcare. In fact, there has been no formal research on ABSA for Norwegian at all, regardless of domain. Although most studies on sentiment analysis within healthcare lean towards traditional methods for polarity classification such as lexicon-based approaches, Naïve Bayes or SVMs (Jelin et al., 2025), there are a few studies which carry out more novel approaches in a healthcare setting, such as employing transformer-based architectures. A recent study conducted by Alkhnbashi et al. (2024) looked into ABSA for patient feedback using BERT-based and generative models. The authors collected a dataset of approximately 15.000 entries from an English medical forum[1] and employed DeBERTa to automatically annotate about 80% of the dataset. The approach which outperformed the others, was few-shot learning using ChatGPT 3.5 turbo. Another study by AlNasser et al. (2024) perform ABSA on a dataset comprising 12400 arabic tweets collected from 14 healthcare organizations in Saudi Arabia. They use GPT-4 in a few-shot setting to identify the most frequent aspects, and end up using five aspects: medical staff, appointments, customer service, emergency services and pricing. The annotation of the aspects and polarities are performed by GPT-4 in this study, and subsequently validated by humans. The authors fine-tune five Arabic-specific versions of BERT, and conduct experiments where one approach is joint, classifying aspects and polarities as one task, and the other is a pipeline approach. In addition, they leverage a joint-model where embeddings from MARBERT (Abdul-Mageed et al., 2021) are fed as features to an SVM. The latter ends up giving the best accuracy.

There are a few points in which our study differs from the above mentioned research. First off, all of our data is human-annotated – specifically by experts on healthcare patient experience data. Secondly, we include a rather large number of fine-grained aspects with extensive coverage of important topics within the special mental healthcare and general practitioner sub-field. Finally, our study in its entirety is based on an interdisciplinary collaboration between the fields of healthcare and natural language processing, which we believe is a very valuable trait.

[1] https://patient.info/

2.2 National Patient Experience Data Collection

The Norwegian Institute of Public Health (NIPH) conducts national patient experience surveys every year, defining Patient Reported Experience Measures (PREMs) and Patient Reported Outcome Measures (PROMs) across different healthcare settings. In addition to closed-ended questions, these surveys include open-ended free-text questions, in which patients are encouraged to describe their experiences with the respective healthcare services. These generate tens of thousands of comments each year, potentially providing valuable insights into health services (Grob et al., 2019; Rivas et al., 2019). However, they are rarely analyzed because this requires a significant amount of time and resources. Currently, NIPH manually analyzes a random sample of 500 comments from each survey and report national-level findings together with quantitative results. For each sampled comment, the polarity is manually recorded (positive, negative, neutral or mixed), and the material is first broadly sorted by theme, and then refined into sub-themes using conventional content analysis (Hsieh and Shannon, 2005; Norman et al., 2024). As a result of this resource-intensive process, thousands of comments remain unused and are excluded from provider-level reports. This exclusion is ethically problematic and limits quality improvement, as providers value and can benefit from such data (Riiskjær et al., 2012; Grob et al., 2019). There is currently therefore a clear need for a more automated analysis of free-text comments in Norwegian patient experience surveys.

3 NorPaC$_{\text{absa}}$

This section provides more details about the creation of the Norwegian Patient Comment corpus for Aspect-Based Sentiment Analysis. As mentioned above, NorPaC$_{\text{absa}}$ contains manual annotations of aspects and polarities for patient comments gathered from surveys by NIPH targeting two distinct domains; specialist mental healthcare (SMH) and general practitioners (GP).

While our dataset cannot be released due to sensitivity constraints, we do release the annotation guidelines, code, models, and detailed descriptions of our experimental setup in our publicly available repository.[2]

NorPaC$_{\text{absa}}$ is annotated with a total of 25 unique aspect categories, where 11 of them are shared between both GP and SMH. See table 3 for an overview. In addition to the actual aspects, we also include a label 'no aspect', indicating that the sentence in question does not contain any aspect. According to Liu (2015), an aspect is defined as being part of the sentiment target, and since the presence of a sentiment target implies the presence of a polar opinion pointed towards it, a neutral sample can be said to not contain any aspects. This can be confusing in cases where a sentence contains an aspect-indicating word, such as 'the doctor's office' in 'I was at

[2] https://github.com/ltgoslo/From_Pain_to_Praise

	Comments	Aspects	Aspects	Tokens
GP	1242	2761	2.2	40.5
SMH	1140	2802	2.5	49.4
$\text{NorPaC}_{\text{absa}}$	2382	5563	2.3	44.8

Table 1: Number of comments and aspects in total, and average number of aspects and average number of tokens (without punctuation) per comment, for each sub-domain of $\text{NorPaC}_{\text{absa}}$.

	Pos		Neg		Neut		Mix	
	#	%	#	%	#	%	#	%
GP	1128	40.9	1306	47.3	64	2.3	263	9.5
SMH	1418	50.6	1101	39.3	28	1.0	255	9.1
$\text{NorPaC}_{\text{absa}}$	2546	45.8	2407	43.3	92	1.7	518	9.3

Table 2: Counts and percentages of aspects per polarity – Positive, Negative, Neutral, and Mixed – row-normalized by domain.

the doctor's office today'. However, since the sentence does not express any polar opinion, 'the doctor's office' is not treated as an aspect. 'No aspect' is therefore treated as the negative class for aspect classification, as well as a neutral label. In addition to neutral, the dataset contains annotations for positive, negative and mixed sentiment. As mentioned earlier, there are a number of different approaches regarding which attributes to include in the ABSA task. In previous research, some choose to include a target expression in addition to the aspect category and the polarity of the opinion (Zhang et al., 2023). In this work we choose to not explicitly annotate the target expression for the task – we believe the combination of the aspect attributes and associated polarity labels adequately represents the patient feedback. By limiting ourselves to these attributes, we also avoid the issue of missing or implied targets. Besides, as our data is anonymized, many noun phrases would be somewhat meaningless if they were included as targets.

Example 1 shows a comment from the GP domain that has been annotated with the aspect 'telephone and digital communication' with negative associated polarity. Worth noticing about the comment is the capital letter style, in addition to five exclamation marks at the end of the phrase. This colloquial tone is a clear characteristic of the dataset.

(1) *VANSKELIG Å OPPNÅ KONTAKT MED FASTLEGEKONTORET!!!!!*

'HARD TO GET IN TOUCH WITH THE GP OFFICE!!!!!'

Labels: `telephone and digital communication:neg`

3.1 Statistics

In this section, we show some of the most essential statistics to gain a better picture of the contents of $\text{NorPaC}_{\text{absa}}$.

Table 1 gives an overview of the number of comments per sub-domain, as well as the total number of annotated aspects, the average number of aspects and tokens per comment. Comparing GP and SMH, we see that SMH on average both have slightly longer comments, as well as more aspects. Worth noting is that the average number of aspects per comment for both domains are a bit more than 2, supporting the fact that we treat the task as multi-label. We would however like to add that the number of aspects per comment span from anywhere between 1 and all the way up to 12 aspects.

Further, table 2 shows a distribution of the aspects marked as positive, negative and mixed in addition to the number of neutral samples. Neutral is the polarity label which clearly has the least coverage in the dataset. This might suggest that most patients who have chosen to answer the survey usually does so because they have a particular negative and/or positive opinion. Moreover, we see that there is quite a big difference in the coverage of mixed as opposed to the positive and negative labels. This distribution also makes sense given that the polarity labels are connected to specific aspects, and not to the utterance as a whole. This naturally decreases the amount of mixed samples.

3.2 Annotation Process

In the following we describe the annotation of aspects resulting in the $\text{NorPaC}_{\text{absa}}$ dataset and provide further details on the inventory of aspect categories and the annotation effort.

Aspect categories When deciding on the precise inventory of aspect categories for annotation, the national PREMs surveys conducted by NIPH served as the starting point for constructing the aspect framework. In particular, the aspect categories were aligned with established quality dimensions of patient-centered care, such as the Picker Principles.[3] An initial set of annotation guidelines were compiled, where each aspect category was described and illustrative examples were provided to ensure consistent annotation. The final set of aspects consist of 6 main categories and 25 unique sub-categories. Out of these, 16 aspects are specific to the sub-domain of general practitioners (GP) and 19 aspects are related to special mental healthcare (SMH), in which 11 of the aspects are shared between the two. The final annotation guidelines will be made openly available for anyone to read.

Annotation The annotation was performed by three health service researchers who have extensive experience with manual analysis of the national PREMs surveys. In an initial annotation effort, the annotators annotated the same set of 40 examples using the first version of the annotation guidelines and inter-annotator agree-

[3] https://picker.org/who-we-are/the-picker-principles-of-person-centred-care/

Aspect		NorPaC$_{absa}$		GP		SMH	
Full name	Short-name	#	%	#	%	#	%
Healthcare providers and staff							
Competence of providers	cp	401	7.2	247	8.9	154	5.5
Information sharing with patients	isp	88	1.6	44	1.6	44	1.6
Language	lang	13	0.2	12	0.4	1	0.1
Patient–provider/staff relationships	ppr	917	16.5	412	14.9	505	18.0
Time Spent with healthcare Professionals	tshp	265	4.8	102	3.7	163	5.8
Organization of health services							
External cooperation with other services	excos	128	2.3	100	3.6	28	1.0
Internal cooperation and communication	incc	30	0.5	10	0.4	20	0.7
Structure and routines	sr	167	3.0	58	2.1	109	3.9
System-level organization of health services	slohs	193	3.5	101	3.7	92	3.3
Duration of treatment and stays	dur	77	1.4	–	–	77	2.8
Access and availability							
Geographical distance to GP office	gd	19	0.3	19	0.7	–	–
Telephone and digital communication	td	181	3.3	181	6.6	–	–
Waiting times in clinic	wtc	54	1.0	54	2.0	–	–
Waiting time for appointment	wtp	121	2.2	121	4.4	–	–
Workload	wol	66	1.2	66	2.4	–	–
Environment and facilities							
Physical and psychosocial environment	ppe	110	2.0	16	0.6	94	3.4
Activities	act	113	2.0	–	–	113	4.0
Interaction with other patients	iop	51	0.9	–	–	51	1.8
Quality of food and meal routines	qfm	90	1.6	–	–	90	3.2
Treatment							
Medication	med	106	1.9	35	1.3	71	2.5
Stability and continuity in treatment	sct	482	8.7	392	14.2	90	3.2
Forced treatment / coercion	ftc	36	0.7	–	–	36	1.3
Uncategorized / Top-level aspects							
Outcome and impact of treatment / stay	oits	319	5.7	–	–	319	11.4
Patient involvement and participation	pip	68	1.2	–	–	68	2.4
General	gen	1376	24.7	727	26.3	649	23.2
No aspect / Neutral	no-asp	92	1.7	64	2.3	28	1.0
Total		**5563**	100.0	**2761**	100.0	**2802**	100.0

Table 3: Overview of aspects, both fine-grained and coarse-grained (i.e. higher-level categories, shown in bold), with counts and percentages per domain. Values of '–' indicate aspects only annotated for one of the domains.

ment was assessed. A training workshop was then held to refine and consolidate the guidelines further. In the workshop, the results from the double annotation were used to discuss disagreements, agree on consolidated annotations for the training examples and refine the annotation guidelines to reflect this agreement. During the main stage of annotation, the three annotators annotated individually across four rounds. Regular meetings were held after each round to discuss difficult examples and clear up any misunderstandings. After the final round of annotation, a subset of 100 comments was annotated by all three annotators to assess the final inter-annotator agreement (IAA). Both IAA subsets (initial stage and final round) contained 50% from the GP domain and 50% from the SMH domain.

A comparison of IAA-scores between the initial pilot annotation and the last round can be seen in table 4. Worth noting, is that the commonly used Cohen's Kappa (Cohen, 1960) to calculate IAA does not inherently support multi-label problems. To work around this, we ended up using a version that assumes correctness only when a sample between two annotators contains *all* the same labels. As a result, this is a rather strict method that does not account for nuances of correctness within each sample, where for instance 2 out of 3 labels are the same between two annotators. Such an example would be counted as incorrect.

	Initial	Final
A1_A2	0.51	0.79
A1_A3	0.48	0.73
A2_A3	0.53	0.89

Table 4: Comparison of the IAA-scores between the test round and the last round. The left column lists annotator id's, where A1_A2 indicates the agreement between annotator 1 and annotator 2, and so on.

As we can see in table 4, there was substantial improvement between the pilot round and last round, highlighting the importance of high quality annotation guidelines and annotator training. On average across the three annotators, the agreement score increased from 0.51 to 0.8. Emphasizing that the number of unique aspects per comment spans anywhere from 1 up to 12, we believe these scores suggest robust annotations.

4 Methodology

With aspects and their polarities annotated on both the sentence- and comment-level, and with a large label-space of aspect categories, only partially overlapping across the two distinct survey domains and with two levels of granularity, there are many possible design

decisions to account for in modeling. Below, we outline the various experimental setups we explore.

For the practical application setting of analyzing patient experiences with healthcare services, the end-goal is providing information on the comment-level. Given that we have annotations on both the sentence- and comment-level, however, we report results for two settings; (*i*) training models to make predictions on the comment-level directly, and (*ii*) making predictions on the sentence-level and then aggregating these to the comment-level before evaluation. We also report results for predicting both fine-grained and coarse-grained aspect categories. Finally, for all setups we compare predicting aspects and polarities *jointly* versus using a *pipeline* approach where the aspects and polarities are predicted in two subsequent stages. In both cases we evaluate results end-to-end, although we also report results for aspect classification in isolation.

As for the pre-trained language models that we fine-tune, we carry out our most extensive experiments as described above using NorBERT3 – a BERT-based model trained on Norwegian text – specifically the large version which has a size of 353 million parameters.[4] Additionally, we perform joint, comment- and sentence-level experiments using NorMistral-11b-thinking – a generative, Mistral-based decoder model with reasoning capabilities which has been continuously pretrained and instruction-tuned on Norwegian data.[5]

We elaborate more on the various setups using NorBERT3$_{\text{large}}$ below.

Pipeline Approach For the pipeline approach, aspect- and polarity classification are performed in two subsequent stages. Below, we first describe how we perform aspect classification, and then turn to how we expand this to include polarities.

Example 2 shows a comment from the GP domain and the aspects it is annotated with. For the aspect classification task, the associated polarity labels are ignored.

(2) *Oppleves som vanskelig å komme igjennom på telefon og få time innen rimelig tid . Resepsjonist kan virke lite hjelpsom i telefon. Lege virker oppriktig intressert i min helse under time, men kunne tatt seg bedre tid.*

'Find it hard to get through by phone and book an appointment within a reasonable time frame. Receptionist can seem unhelpful on the phone. GP seems genuinely interested in my health during appointment, but could have spent more time.'

Labels: `waiting time for appointment`, `patient-provider relationship`, `telephone and digital communication`, `time spent with healthcare professionals`

[4] https://huggingface.co/ltg/norbert3-large

[5] https://huggingface.co/norallm/normistral-11b-thinking

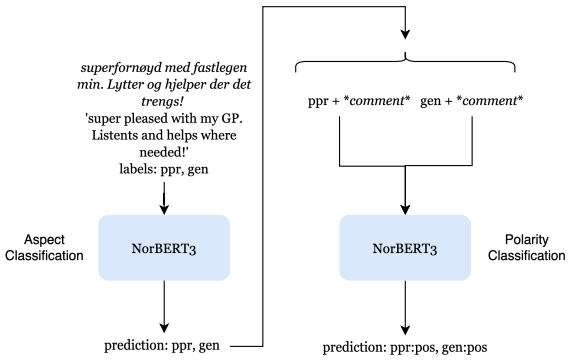

Figure 1: Illustration of the pipeline approach, where a comment from the GP domain goes from aspect- to polarity classification.

In the next step, the text and the predicted aspects are given as a concatenated input to the NorBERT3$_{\text{large}}$ model. The aspects are provided with their full name, as shown in table 3. To be able to connect polarity predictions to specific aspects, only one aspect is given as input together with the text at a time, meaning that the same text may be passed to the polarity classifier several times, to take account of multiple detected aspects. Figure 1 shows an example where a comment from the GP domain is first classified with multiple aspects and then passed to the polarity model.

Joint Approach For the joint approach, we classify aspects and associated polarities in one step. This approach requires a pre-processing step of the labels, where all aspects and polarity labels are simply merged together to form a single label. For instance, if a comment is labeled with the aspect 'Medication' and associated polarity, 'positive', the final label will be 'medication_positive'. This also means that the number of unique labels will be much higher than the amount of unique aspects alone. For the fine-grained aspects setup, this step results in a total of 55 unique labels from GP and SMH combined. Although the method is simple, the classification task becomes quite complex. Since there are so many different labels, there are also fewer samples marked with each of them, making evaluation a bit more challenging.

5 Encoder Model

In this section, we conduct all our experiments using NorBERT3$_{\text{large}}$. As mentioned earlier, there are several options when it comes to choosing domains for training and testing. As our initial experiments have shown a clear advantage of training on the full NorPaC$_{\text{absa}}$ dataset when testing on both domains of the dataset, this is what we report for in the upcoming result sections.

5.1 Results & Discussion

Aspect Classification We start our experiments with aspect classification only, as this is the most complex part of the task, and is what creates a foundation for

		Aspect only		Joint	
		GP	SMH	GP	SMH
Fine-grained	Comment	$67.06^{\pm 0.95}$	$69.78^{\pm 2.83}$	$64.38^{\pm 0.61}$	$62.65^{\pm 1.14}$
	Aggregated	$66.41^{\pm 1.26}$	$72.72^{\pm 1.40}$	$65.95^{\pm 1.93}$	$70.29^{\pm 1.92}$
Coarse-grained	Comment	$71.87^{\pm 2.00}$	$72.81^{\pm 1.37}$	$68.84^{\pm 1.90}$	$68.80^{\pm 1.03}$
	Aggregated	$74.60^{\pm 0.75}$	$76.49^{\pm 1.69}$	$75.51^{\pm 1.22}$	$74.81^{\pm 1.12}$

Table 5: Weighted average F1 for aspect classification averaged across five seeds. For the joint results, the aspects are extracted from the aspect+polarity labels.

the subsequent polarity classification. We here mainly compare two approaches; the first constitutes the first step in the pipeline approach where aspects are classified in isolation, before being provided as input to a second stage of polarity classification for each aspect. Hence, for the aspect classification comparison, the training data here only consists of the gold aspects, and not the associated polarity. The other approach which we compare against, is the joint method, where the training data consists of gold aspects that are concatenated with gold polarities, as described in section 4. Since we are only interested in the aspect classification in this section, we filter away the polarity part of the labels during evaluation. The main research question to be answered in this comparison will therefore be whether it is advantageous to include both aspect and polarity, or aspect information only when classifying aspects.

Table 5 reports the performance on classifying aspects only for both the model trained solely on aspects, and the model trained on concatenated aspect-polarity pairs. Looking at the table, we see that for every test environment except for one, it is better to train on aspects only rather than the concatenation of aspect and polarity. The only environment where the joint model beats the aspect-only model, is when the sentence-level model is trained on coarse-grained aspects and evaluated with sentence-predictions aggregated to the comment-level for the GP domain. For the aspect-only model there is also a quite clear trend of SMH seeing higher scores then GP, while this trend is less clear, or even the other way around, for the joint model. Our earlier experiences with sentiment analysis classification only with this dataset, however, shows the opposite trend. This leads us to believe that the reason for SMH showing better performance this time is due to sensitivity to the number of samples per class in the test set, and not necessarily due to the language differences between the domains. (Rønningstad et al., 2025). This also becomes more clear as we have seen that this pattern often shows up for the test set, but not as often for the validation set.

When it comes to text granularity, we observe that it is mostly advantageous to train the model on sentence-level data before evaluating on comment-aggregated predictions, compared to training the model on comment-level data directly. One of the reasons for this outcome, could be that this splitting of comments simply leads to more training samples. Another potential explanation, could be that shorter texts have a clearer relationship with their annotated aspects. Besides – there will also be fewer aspects associated with each sentence compared to with each comment. All in all, it seems to be beneficial to use an aspect-only trained model to classify aspects at the sentence-level.

On further analysis of the output of the models, we observed several cases where the model's predictions strictly were incorrect compared to the annotations, yet seemed reasonably justified when examining the actual content of the comment. It is important to note that for many of the comments, there is not necessarily one correct answer regarding which aspects are relevant. This is a fact that was agreed upon between the annotators, which is also reflected in some of the model classifications. Such an example can be seen in example 3.

(3) *Skulle ønske at personalet hadde vært mere i stuen på kveldstid*

'I wish that the staff would have been more in the living room during the afternoon'

true: `time spent with healthcare professionals`
pred: `time spent with healthcare professionals, physical and psychosocial environment`

Example 3 shows a comment from the SMH domain where the patient wishes that the staff would have been more in the living room during the afternoons. The true label for this sample is 'time spent with healthcare professionals'. The model, which is the one trained on comment-level and fine-grained aspects in this case, has additionally assigned 'physical and psychosocial environment'. Considering the content of the feedback, it is not unreasonable to think that absence of staff in a social area is something that affects the psychosocial environment.

Aspect Imbalance Although a label set with extensive coverage mostly comes with positive traits, it also brings about the challenge where introduction of more fine-grained aspects leads to a decreased number of samples per aspect. This is especially visible for some of the labels with few total instances, which leads to some of them not being present in the randomized test split, as seen in table 6. In addition, our task is multi-label, which restricts an even distribution of labels across splits even more as altering the distribution of one aspect cannot be done without affecting another. To cope with this challenge, we perform cross-validation across five folds applied to the full $\text{NorPaC}_{\text{absa}}$ dataset. This ensures that each aspect is present in at least one of the test folds. This approach will further be closer to the practical application of the task, as the final model applied on new data will have undergone training on the full dataset. We further investigate the more frequently occurring labels by including evaluation of all fine-grained labels that

occur five or more times in the test split. As most of the coarse-grained labels already exceed this frequency, they are left out from this evaluation.

By looking at the cross-validation results in table 6, we see that all scores that were previously zero, have increased. Moreover, when comparing the overall weighted average F1-scores between cross-validation and the evaluation of the original test split, we can tell that the scores of the GP domain are very close to each other, while the scores of the SMH domain are a bit further apart. This may confirm our earlier discussion on the SMH domain's potential sensitivity to sample frequency.

Looking at the scores for the labels occurring five or more times in the test set, we see that all scores have either increased or stayed the same compared to the evaluation of the full label set. The highest macro F1-score reaches 93.05%, and no scores go below 67.86%. Further, the overall weighted average F1-scores have increased from 66.41% to 77.36% for the GP domain, and from 72.72% to 83.05%, showing solid performance in the most prominent aspects of NorPaC$_{absa}$.

End-to-End Classification We proceed to look at evaluation of the full task – that is, classification of both aspect and polarity, where we compare the performance between the two-stage pipeline approach and the one-stage joint approach. As we began with evaluating aspect classification only, a natural next step would be to evaluate polarity only. However, as the polarity labels in this case are highly dependent on the aspects, this is not an easy task. For the pipeline approach, this is unproblematic as polarity can be predicted on gold aspects. This is however not the case with the joint approach, as we would have to remove the samples where the aspect is incorrect. This is because those predictions have no gold label to compare with. For these reasons, we evaluate the full task of aspect and polarity classification in this section.

Earlier in table 5, we saw that the aspect model outperformed the joint model when classifying aspects. Looking at the performance for the full task in table 7, the joint model seems to have caught up with the aspect model – at least at the aggregated level, for both the fine-grained and coarse-grained labels. Comparing the two approaches, the joint models seem to benefit much more from sentence-level training than the pipeline models. In contrast, the pipeline performance decreases when trained on sentences.

Although polarity evaluation is an unrealistic task in this case, we have made observations in the output for both setups that indicate that the mixed label in particular is challenging compared to the other polarity labels. This is not surprising as the number of mixed labels only cover 9.3% of NorPaC$_{absa}$, as discussed in section 3.1.

All in all, the joint approach outperforms the pipeline setup when it comes to modeling the full task, and seems especially advantageous at the aggregated text level.

6 Decoder Model

In this section, we conduct our experiments with the Mistral-based NorMistral-11b-thinking,[6] an instruction-tuned 11 billion parameter model trained on open Norwegian texts. As implied by the name, this model has 'thinking' capabilities, meaning that it has been fine-tuned to break problems into smaller steps prior to generating the final output (Samuel et al., 2025).

We conduct zero-shot- and few-shot experiments, as well as fine-tuning the model. Like masked models, there are several possible approaches to fine-tuning of a decoder model for ABSA. Since this is a generative model, we decide to make the most of its capabilities, and therefore classify aspect and polarity in one step by prompting it do so. We also perform fine-tuning in a multi-label setting. Fine-tuning is performed for both comment- and sentence-level, including for fine-grained and coarse-grained aspects. Hence, these results will be comparable to the joint model in a full task setting, as reported in table 7.

For zero-shot, few-shot and fine-tuning, we use the same prompt constructed with the intention of defining the task in a clear manner, enabling aspect- and polarity classification in one step. This involves clarifying the possible aspects and sentiment labels to be assigned, as well as the format to give the answer in. The full prompt can be seen in example 4 under appendix B.

6.1 Results & Discussion

In this section, we proceed to discuss the results from NorMistral in the three settings. Table 8 shows weighted average F1-scores for zero-shot and few-shot settings, as well as fine-tuned versions of the model.

Zero-shot As we can see in table 8, the performance of this task in a zero-shot setting is generally low. We find that the models using coarse-grained aspects perform better than using fine-grained aspects, especially for the SMH domain. When it comes to the text granularity, comment-level yields better performance than sentence-level in all cases.

On further analysis of the output, one of the biggest issues is likely that the model comes up with labels that are not part of the pre-defined set of aspect categories. This can for instance be keywords mentioned in the comment itself. We further observe that the model does not fully adhere to the requested json-format, and sometimes not to the format of aspect-sentiment pairs either. Specifically, the rule introduced in the prompt stating that the 'neutral' label should not be used with any other labels than the 'no aspect' label, is broken repeatedly during generation. While most of the output is quite close to a correct json-format, there are often minor issues that make extraction problematic. In addition, there are cases where the predicted set of labels seem to have no connection to the associated text, almost making it

[6] https://huggingface.co/norallm/normistral-11b-thinking

Aspect	GP				SMH			
	Test F1	Instances	F1 5-fold	F1 # > 5	Test F1	Instances	F1 5-fold	F1 # > 5
Healthcare providers and staff	84.41	43	76.52		80.58	50	76.92	
Competence of providers	65.05	18	58.94	75.8	84.87	11	60.26	93.05
Information sharing with patients	0.00	1	36.22	*	96.00	2	47.48	*
Language	66.67	1	47.78	*	0.00	0	100.0	*
Patient–provider/staff relationships	64.69	26	73.01	73.74	73.18	39	76.68	85.63
Time spent with healthcare professionals	71.55	6	70.22	78.36	58.88	6	58.59	72.0
Organization of health services	61.73	9	59.53		56.23	13	53.65	
External cooperation with other services	85.71	3	77.88	*	0.00	0	27.95	*
Internal cooperation and communication	0.00	0	29.17	*	0.00	0	36.46	*
Structure and routines	10.0	1	32.28	*	0.00	2	44.87	*
System-level organization of health services	52.04	5	48.12	67.86	68.10	4	53.39	*
Duration of treatment and stays	–	–	–	–	–	7	68.57	71.55
Access and availability	75.7	8	81.15		–	–	–	–
Geographic distance to GP office	0.00	0	48.35	*	–	–	–	–
Telephone and digital communication	68.25	4	78.11	*	–	–	–	–
Waiting times in clinic	73.33	1	83.02	*	–	–	–	–
Waiting time for appointment	73.90	3	75.56	*	–	–	–	–
Workload	93.33	1	69.39	*	–	–	–	–
Environment and facilities	0.00	1	48.89		70.96	18	76.68	
Physical and psychosocial environment	0.00	1	62.63	*	37.33	4	61.32	*
Activities	–	–	–	–	74.80	9	74.16	79.79
Interaction with other patients	–	–	–	–	37.33	3	50.97	*
Quality of food and meal routines	–	–	–	–	86.11	5	84.57	86.11
Treatment	71.07	25	66.34		79.61	7	64.49	
Medication	96.00	2	65.14	*	90.56	5	75.88	90.56
Stability and continuity in treatment	68.05	23	65.74	78.78	64.00	3	50.49	*
Forced treatment / coercion	–	–	–	–	50.00	1	71.35	*
Uncategorized / Top-level aspects								
Outcome and impact of treatment/stay	–	–	–	–	64.30	18	58.74	77.4
Patient involvement and participation	–	–	–	–	57.14	5	46.93	57.14
General	67.29	50	62.22	80.88	75.53	58	60.5	85.36
No aspect / Neutral	62.93	7	62.89	74.77	73.75	9	55.66	84.11
Fine-grained weighted avg. F1	$66.41^{\pm 1.26}$		$65.27^{\pm 0.77}$	$77.36^{\pm 0.96}$	$72.72^{\pm 1.40}$		$62.63^{\pm 1.71}$	$83.05^{\pm 1.19}$
Coarse-grained weighted avg. F1	$74.60^{\pm 0.75}$		$68.45^{\pm 1.88}$		$76.49^{\pm 1.69}$		$64.92^{\pm 1.11}$	

Table 6: Average macro F1-scores per label in three settings; averaged across five seeds evaluated on the test set, averaged across five folds in cross validation, and evaluated across five seeds for the fine-grained labels with five or more instances in the test set. All scores are reported for the sentence-trained NorBERT3$_{\text{large}}$. The 'Instances' columns show the number of instances per label in the test split. Values of '–' indicate aspects only annotated for one of the domains, and values of '*' indicate labels with fewer than five instances in the test set.

		Pipeline – Full Task		Joint – Full Task	
		GP	SMH	GP	SMH
Fine-grained	comment	$60.01^{\pm 1.13}$	$67.77^{\pm 0.98}$	$60.75^{\pm 1.07}$	$58.34^{\pm 0.74}$
	aggregated	$57.17^{\pm 0.87}$	$64.86^{\pm 0.87}$	$60.76^{\pm 1.84}$	$66.79^{\pm 1.66}$
Coarse-grained	comment	$65.62^{\pm 0.42}$	$67.55^{\pm 0.92}$	$65.11^{\pm 1.94}$	$64.44^{\pm 0.79}$
	aggregated	$64.99^{\pm 0.6}$	$66.56^{\pm 0.82}$	$69.99^{\pm 1.43}$	$70.35^{\pm 1.25}$

Table 7: Weighted average F1-scores for the full task averaged across five seeds. For the pipeline approach, the final concatenated aspect+polarity labels from respectively the first model and the second model are evaluated. For the joint approach, the aspect+polarity labels given as a single-stage task are evaluated.

look like they were chosen by chance. However, we do see that the reasoning traces in the model's output reflect the instructions correctly, although this does not always apply to the final answer.

Few-shot Moving on to our few-shot experiments, we use the prompt as discussed, but provide it with two random examples from the dataset. These include the patient feedback itself, and the associated labels. Before landing on a 2-shot setting, higher shot-settings were attempted. As there was no consistent gain in performance for the two sub-domains with the increasing number of shots, 2-shot became the final choice.

Judging from the results in table 8, we can see that the scores have increased somewhat for some of the settings compared zero-shot. This is especially the case for the GP domain, in particular when looking at the fine-grained level. For SMH, there is a slight increase in all settings, except for the aggregated, fine-grained level, which has a small drop in performance.

When it comes to the effects of including sentence-level feedback vs. comment-level feedback, we see the same pattern that we saw for zero-shot – including comment-level text yields better performance than sentence-level. We assume that giving access to full comments provides better context for the model.

			GP	SMH
Zero-shot	Fine-grained	comment	27.92	32.31
		aggregated	20.87	28.8
	Coarse-grained	comment	28.66	42.91
		aggregated	28.63	37.08
Few-shot	Fine-grained	comment	35.96	34.84
		aggregated	26.65	27.22
	Coarse-grained	comment	32.62	43.86
		aggregated	30.1	37.91
Fine-tuned	Fine-grained	comment	$36.66^{\pm 2.37}$	$41.7^{\pm 1.12}$
		aggregated	$47.13^{\pm 1.93}$	$52.93^{\pm 2.5}$
	Coarse-grained	comment	$32.55^{\pm 1.4}$	$39.23^{\pm 0.99}$
		aggregated	$26.12^{\pm 1.86}$	$41.33^{\pm 3.56}$

Table 8: Weighted avg. F1-scores for zero-shot, few-shot and a fine-tuned version of NorMistral-11B-thinking, all using the prompt provided in the appendix.

When inspecting the model output, we observe some of the same tendencies as for the zero-shot outputs. First, there are still issues with the json-format as well as some of the aspect-sentiment pairs. We also see a few occurrences where the model repeats labels within the same sample, although this error is filtered out when calculating the scores. Further, there are cases where the predicted labels do not make sense when looking at the given text. It may be that some of the predictions that do not connect logically to the given text, may be inspired by the provided few-shot examples. However, this does not seem to be a consistent pattern. Apart from this, there seems to be a slight improvement when it comes to hallucination of labels that are not part of the pre-defined label set. Although this error is still present, we believe that this may be the main reason for the slight improvement in performance compared to zero-shot.

Fine-tuning As for the fine-tuning, we use the same prompt that we have used for zero-shot and few-shot. For each training sample, we concatenate the prompt and the comment text, serving as the text to classify, along with the gold labels given as the assistant response.

Taking a look at the results of the fine-tuned version in table 8, we see that the results are somewhat inconsistent. Within the GP domain, we see that the fine-grained results have improved compared to zero-shot and few-shot settings. This is especially the case for the fine-grained, aggregated level, which has a clear advantage of fine-tuning. Looking at the coarse-grained level, this advantage is less clear. While the comment-level out-performs the zero-shot setting, it shows marginally lower performance compared to few-shot. Surprisingly, the aggregated sentence-level yields the lowest performance when in a fine-tuned setting. Moving on to the SMH domain, we see an overall increase at the fine-grained level, especially for the aggregated level, as we have seen for GP. This is, however, not the case for the coarse-grained labels. At comment-level, we see that fine-tuning reveals lower performance than both zero- and few-shot settings. In contrast, we observe that the aggregated level benefits from fine-tuning.

Upon further analysis of the fine-tuned models, the output almost exclusively follow the given json-format given by the prompt as well as the training samples. We also have not observed hallucinated aspects in the outputs. However, there are still some cases where the neutral sentiment is put together with other labels than the 'no aspect' label. Further, we observe that there is a tendency where the fine-tuned models over-represent the most frequent labels of our dataset. In particular, this applies to the 'general' and the 'patient-provider/staff relationships' aspects. As a consequence, the other, perhaps more specialized labels, are under-represented.

7 Summary

This paper has reported experimental results for aspect-based sentiment analysis, trained and evaluated on $\text{NorPaC}_{\text{absa}}$– a dataset comprising 2382 free-text patient comments annotated by experts on healthcare experience data. Covering a range of relevant aspects within patient experience, our dataset enables a solid foundation for training domain-targeted sentiment analysis models. In turn, this can assist healthcare researchers in the analysis of patient feedback, opening for a much more efficient approach compared to today's manual analysis. This means a much larger volume of patient feedback can be analyzed, potentially leading to improved patient satisfaction and clinical outcomes.

We have reported experimental results for fine-tuning both encoder- and decoder models on our data, and compared them for several levels of text- and label-granularities. Our findings show that a considerably smaller encoder model, specifically $\text{NorBERT3}_{\text{large}}$, is a better choice compared to the decoder model, NorMistral-11b-thinking. Further, training on sentence-level data where the predictions are later aggregated to the comment-level, is shown to be beneficial compared to training on comment-level annotations directly. While coarse-grained aspects naturally lead to higher performance scores than fine-grained aspects, as the number of unique labels decrease, results for fine-grained aspects do not fall too far behind. Finally, modeling the task using a joint approach where concatenated aspect and polarity labels are classified in a single step proves to be feasible, and maybe even preferable, despite having a notably larger set of labels.

Limitations

While one of our experimental setups included the decoder-based NorMistral-11b-thinking, we acknowledge that prompt construction is something that represents a broad landscape of design choices. Given time constraints and the already extensiveness of the initial encoder-based setup using $\text{NorBERT3}_{\text{large}}$, we did not create a systematic setup to explore different prompt strategies.

We would further like to mention the 'general' aspect, which is a label that covers the examples where the content is too general to be pin-pointed to one of the other pre-defined aspects. While general satisfaction or dissatisfaction is valuable information, this type of feedback is very frequent in our dataset, and may lead to over-representation of the category in the outputs.

Ethical Considerations

As our dataset contains patient experiences that may be of personal manner, the content has to be handled with care. To ensure the privacy of patients, the data has undergone an anonymization procedure at NIPH before transferal to UiO, but will for the same reason not be made public. Because the data cannot be published, we have restricted our modeling experiments to only use open source models that can be run locally under controlled environments.

Acknowledgements

We would like to thank Elma Jelin, Mona Haugum and Inger Opedal Paulsrud for their annotation contributions. This work was supported by two research projects funded by the Research Council of Norway (RCN), namely 'Strengthening the patient voice in health service evaluation: Machine learning on free-text comments from surveys and online sources', funded by a HELSEVEL grant from RCN (project no. 331770). Moreover, the computations were performed on resources provided through Sigma2 – the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway, as well as Fox – a High Performance Computing cluster for Educloud Research users.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Omer S. Alkhnbashi, Rasheed Mohammad, and Mohammad Hammoudeh. 2024. Aspect-Based Sentiment Analysis of Patient Feedback Using Large Language Models. *Big Data and Cognitive Computing*, 8(12):167.

Seba Alnasser and Sarab Almuhaideb. 2024. Listening to Patients: Advanced Arabic Aspect-Based Sentiment Analysis Using Transformer Models Towards Better Healthcare. *Big Data and Cognitive Computing*, 8:156.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Ganpat Singh Chauhan, Ravi Nahta, Yogesh Kumar Meena, and Dinesh Gopalani. 2023. Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, 49:100576.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46. _eprint: https://doi.org/10.1177/001316446002000104.

Rachel Grob, Mark Schlesinger, Lacey Rose Barre, Naomi Bardach, Tara Lagu, Dale Shaller, Andrew M. Parker, Steven C. Martino, Melissa L. Finucane, Jennifer L. Cerully, and Alina Palimaru. 2019. What Words Convey: The Potential for Patient Narratives to Inform Quality Improvement. *The Milbank Quarterly*, 97(1):176–227.

Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288.

Elma Jelin, Lilja Charlotte Storset, Rebecka M Norman, Hilde Hestad Hestad Iversen, Lina Harvold Ellingsen-Dalskau, Petter Mæhlum, Erik Velldal, Lilja Øvrelid, and Oyvind Bjertnaes. 2025. From words to action? A scoping review on automatic sentiment analysis of patient experience comments from online sources and surveys. *BMJ Health & Care Informatics*, 32(1):e101631.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge.

Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid, and Erik Velldal. 2024. It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 8–19, Torino, Italia. ELRA and ICCL.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Rebecka Maria Norman, Elma Jelin, and Oyvind Bjertnaes. 2024. Multimorbidity and patient experience with general practice: A national cross-sectional survey in Norway. *BMC Primary Care*, 25(1):249.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan

Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Erik Riiskjær, Jette Ammentorp, and Poul-Erik Kofoed. 2012. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, 24(5):509–516.

Carol Rivas, Daria Tkacz, Laurence Antao, Emmanouil Mentzakis, Margaret Gordon, Sydney Anstee, and Richard Giordano. 2019. *Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study*. Health Services and Delivery Research. NIHR Journals Library, Southampton (UK).

Egil Rønningstad, Lilja Charlotte Storset, Petter Mæhlum, Lilja Øvrelid, and Erik Velldal. 2025. Mixed Feelings: Cross-Domain Sentiment Classification of Patient Feedback. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 537–543, Tallinn, Estonia. University of Tartu Library.

David Samuel, Lilja Øvrelid, Erik Velldal, and Andrey Kutuzov. 2025. Fluent Alignment with Disfluent Judges: Post-training for Lower-resource Languages. In *Proceedings of the Fourteenth International Conference of Learning Representations (ICLR 2026)*, Rio de Janeiro, Brazil.

Paul F. Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *arXiv preprint*. ArXiv:2310.18025 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

A Model Hyperparameters

A.1 NorBERT3$_{large}$ for fine-grained labels

Parameter	Value
epochs	10
batch size	8
learning rate	5e-05
warmup ratio	0.0
weight decay	0.01
hidden dropout	0.1
Seeds	[42, 43, 44, 45, 46]

Table 9: Hyperparameters for NorBERT3$_{large}$. The contents of the table applies to all models trained on fine-grained aspects, including the sequential and joint approach for ABSA.

A.2 NorBERT3$_{large}$ for coarse-grained labels

epochs	10
batch size	8
learning rate	3e-05
warmup ratio	0.0
weight decay	0.01
hidden dropout	0.1
Seeds	[42, 43, 44, 45, 46]

Table 10: Hyperparameters for NorBERT3$_{large}$. The contents of the table applies to all models trained on coarse-grained aspects, including the sequential and joint approach for ABSA.

A.3 NorMistral-11b-thinking

LoRA Rank	8
LoRA α	16
Quantization	4-bit
Target modules	[q, v]
Max sequence length	1024
Epochs	4
Batch size	4
Seeds	[42, 43, 44, 45, 46]

Table 11: The hyperparameters used for fine-tuning NorMistral-11b-thinking.

Hyperparameters for fine-tuning

B Prompt

The following prompt in example 4 was used for the zero-shot and few-shot experiments, as well as for the fine-tuning setup.

(4) `system`: Du er en ekspert på analyse av pasienttilbakemeldinger. Tildel ett eller flere aspekter og deres tilhørende sentiment til den gitte pasienttilbakemeldingen. Svar kun på json-format der *aspekt* byttes ut med det faktiske aspektet og *sentiment* med det tilhørende sentimentet på denne formen: [{*aspekt*:*sentiment*}, ..] eller dersom tilbakemeldingen er nøytral: [{'ingen aspekt':'nøytral'}]. Hvis pasienttilbakemeldingen inneholder et generelt utsagn som ikke kan knyttes til noen av de andre aspektene, benyttes aspektet 'generelt' med tilhørende sentiment.
Mulige aspekter: *aspektene i deres fulle navn*.
Mulige sentiment: positiv, negativ, blandet, nøytral (brukes kun sammen med 'ingen aspekt').
`user`: Tilbakemelding: *tekst*

'`system`: You are an expert in analyzing patient feedback. Assign one or more aspects and their associated sentiment labels to the given patient feedback. Answer in json format only where *aspect* is replaced with the actual aspect and *sentiment* with the associated sentiment on this format: [{*aspect*:*sentiment*}, ..] or, if the feedback is neutral: [{'no aspect':'neutral'}]. If the feedback contains a general statement that cannot be pin-pointed to any of the other aspects, the 'general' aspect is to be used.
Possible aspects: *the aspects in their full names*.
Possible sentiment: positive, negative, mixed, neutral (only to be used with 'no aspect').
`user`: Feedback: *text*'