It's Difficult to be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments

Petter Mæhlum[1], David Samuel[1], Rebecka Maria Norman[2], Elma Jelin[2], Øyvind Andresen Bjertnæs[2], Lilja Øvrelid[1], Erik Velldal[1]
[1]University of Oslo, [2]Norwegian Institute of Public Health,
[1]{pettemae, davisamu, liljao, erikve}@ifi.uio.no,
[2]{rebecka.norman, elma.jelin,oyvindandresen.bjertnaes}@fhi.no,

Abstract

Sentiment analysis is an important tool for aggregating patient voices, in order to provide targeted improvements in healthcare services. A prerequisite for this is the availability of in-domain data annotated for sentiment. This article documents an effort to add sentiment annotations to free-text comments in patient surveys collected by the Norwegian Institute of Public Health (NIPH). However, annotation can be a time-consuming and resource-intensive process, particularly when it requires domain expertise. We therefore also evaluate a possible alternative to human annotation, using large language models (LLMs) as annotators. We perform an extensive evaluation of the approach for two openly available pretrained LLMs for Norwegian, experimenting with different configurations of prompts and in-context learning, comparing their performance to human annotators. We find that even for zero-shot runs, models perform well above the baseline for binary sentiment, but still cannot compete with human annotators on the full dataset.

Keywords: patient feedback, sentiment analysis, generative models, in-context learning, annotation

1. Introduction

The Norwegian government has a long tradition of collecting data on patient experiences in the form of surveys, and recently this has also included of unstructured free-text comments. The application of sentiment analysis (SA) to these texts is expected to provide valuable information on patient experiences, which can then be used to improve care at both district and national levels.

This paper documents a large-scale annotation effort to add comment- and sentence-level polarity to patient feedback, representing a collaboration between NLP researchers and health professionals. Specifically, we target patient comments on experiences with General Practitioners and Special Mental Healthcare. In addition to presenting the annotation guidelines and an analysis of the resulting dataset, we also include experimental results on augmenting the human annotations with predictions by pretrained large language models (LLMs). Using two recently released generative LLMs for Norwegian based on the T5 and Mistral architectures (Raffel et al., 2020; Jiang et al., 2023), we present results for different prompts combined with zero- and few-shot learning. We also compare and discuss the differences in error types made by human annotators and the models.

Due to privacy concerns and the sensitive nature of patient feedback, the underlying text material can unfortunately not be openly distributed, but we publish the prompts, the class distributions and the annotation guidelines.[1]

[1]https://github.com/ltgoslo/Sentiment-Annotation-of-Patient-Comments/

2. Background and Motivation

Importance of patient feedback Systematic reviews of the literature show that positive patient experiences are associated with better patient safety, better effectiveness, higher levels of adherence, and lower healthcare utilization (Anhang Price et al., 2014; Doyle et al., 2013).

An important patient-oriented tool at the national level is the national system for measuring patient experiences. The purpose of the system is to systematically measure patient experiences with health services, as a basis for quality improvement, management of health services, free patient choice, and public accountability. To underpin these goals, quantitative results from surveys are produced and presented at different health care levels, e.g., results for hospitals, for health regions, and results at the national level.

NIPH Surveys The Norwegian Institute of Public Health (NIPH) has conducted many national patient experience surveys. All surveys include one or more open-ended questions in which patients are encouraged to write about their experiences with the health service, which is equivalent to tens of thousands of comments each year. These comments present a rich data source on health service evaluation (Grob et al., 2019b; Rivas et al., 2019), but are mostly unused due to the time and resources needed to analyze them. NIPH's current approach is to conduct manual content analysis of a random sample of 500 comments in each survey and report main findings at the national level alongside quantitative results. Furthermore, providers

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 8–19
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

might get access to the data for their patients, but most providers lack the competence, systems and resources to analyze qualitative data. This means thousands of free text comments from each survey are excluded from further analysis and consequently also from provider-level reports. This exclusion is problematic from an ethical point of view, but also because these types of data at lower levels are highly valued by providers (Riiskjær et al., 2012) and are well-suited for use in quality improvement (Grob et al., 2019b). Free-text comments from surveys have the potential to nuance the quantitative data. For instance, a substantial proportion of patients with the highest quantitative scores describe negative experiences in free-text comments (Iversen et al., 2014), indicating that these could be used for differentiating patients at the higher end of the scale. Thus, one can expect added value of quantitative indicators at the provider-level based on the qualitative feedback of text comments.

Therefore, there is a clear need for an innovative and highly efficient method for analyzing large amounts of patient comments. In this paper we describe the first steps towards the automatic analyses of Norwegian free-text comments from patients. Feedback reports with results of free text analysis at the provider-level will make them more relevant and actionable for clinicians and managers who want to improve quality (Grob et al., 2019b; Riiskjær et al., 2012), thus possibly also strengthening the patient's voice in quality improvement. Alleviating the workload and costs associated with annotating data for these systems constitutes an important step in this direction.

3. Previous work

NLP for Patient Feedback Khanbhai et al. (2021) present a systematic review on the application of NLP and machine learning techniques to patient experience feedback. It shows that 80% of the surveyed studies applied language analysis techniques to patient feedback from social media sites followed by structured surveys. These studies include work based on both supervised and unsupervised learning for text and sentiment analysis (SA). To provide data for supervised SA, previous work relies heavily on manual classification of a subset of data by themes and sentiment (Alemi et al., 2012; Bahja and Lycett, 2016; Doing-Harris et al., 2017; Greaves et al., 2013; Hawkins et al., 2016; Huppertz and Otto, 2018; Wagland et al., 2016; Jiménez Zafra et al., 2017). In previous work, comments extracted from social media have also been analyzed using an unsupervised approach; however, free-text comments from surveys are typically analyzed using supervised machine learning (Khanbhai et al., 2021). Khanbhai et al. (2021) discusses that comprehensive reading of all comments within the dataset remains the 'gold standard' method for analyzing free-text comments, and that this is currently the only way to ensure that all relevant comments are coded and analyzed, demonstrating that language analysis using an ML approach is only as good as the dataset used to inform it. Other studies recently published in this area are all examples of how NLP and SA can be used to make the information more accessible and usable in various quality improvement initiatives, for example, using dashboards, pipelines, and visualization (Alexander et al., 2022; Khanbhai et al., 2022; Rohde et al., 2022; van Buchem et al., 2022).

Norwegian Sentiment Analysis To our best knowledge, there has been no previous work on sentiment analysis (SA) for free-text patient feedback, or for any user-generated text in Norwegian. The bulk of previous work on Norwegian SA in general relies on the NoReC dataset of multi-domain reviews collected from various news sources (Velldal et al., 2018). Based on an annotated subset, Øvrelid et al. (2020) have published a fine-grained SA corpus (NoReC$_{fine}$) along with annotation guidelines, on which we partly base our manual annotation effort, further described in Section 4. NoReC$_{fine}$ contains annotations for fine-grained sentiment: annotating the opinion holder, target and polarity. In addition to this, previous work has focused on entity-level aggregation of SA annotation (Rønningstad et al., 2022), and also on improving existing models using data augmentation based on a masked language model (Kolesnichenko et al., 2023). It's worth noting that, being written by professionals reviewers, the documents in NoReC do not contain many of the features that are typical of user-generated data, as in our patient comments, apart from the obvious differences in domain.

Zero/Few-Shot Evaluation In this work, the main focus will be on decoder-based, or generative, language models and the evaluation of their capabilities as sentiment annotators. These types of models have been shown to perform well in zero-shot or few-shot settings (Brown et al., 2020; Wei et al., 2022) with limited annotated data. Given the possibility of accessing these models via natural language prompts, they are arguably also easier to use than traditional models, especially in consideration of health professionals who may not have programming experience. We further focus exclusively on openly available models that can be run locally and do not risk leaking of sensitive data via an API to a proprietary service. This is crucial in our data setting, where surveys are considered sensitive data.

4. Annotation

Scope and Sources Our data comes from free-text comments from the NIPH patient experience surveys described above. The surveys cover various different domains, but we focus on two subsets of data: experiences with General Practitioners (GPs) and evaluation of Special Mental Healthcare (SMH). The data from these surveys has been the focus of earlier research (Kjøllesdal et al., 2020; Iversen et al., 2022), but the use of machine learning methods on the free-text comments is new. While the domain (health-related) and the genre (user-generated text) in these surveys are quite different from the professional reviews found in the existing NoReC corpus, we adopt a similar annotation setup as that used in the NoReC$_{fine}$ (Øvrelid et al., 2020) annotation effort. The original annotations were done at both comment- and sentence-level, with a three-way intensity scale, together with positive and negative sentiment, as well as explicit mention of neutral sentences, indicating that there is no expression of sentiment. Rather than using the full space of distinct labels allowed by this annotation scheme, in the experiments reported later we will only use four, corresponding to positive, negative, mixed, and neutral, ignoring the intensities.

Annotators Annotation was performed by seven researchers in the health service. A set of annotation guidelines was devised based on NoReC$_{fine}$, with some adaptations to the active domain. Certain aspects of NoReC$_{fine}$, such as the distinction between various types of sentiment, and the precise delimitation of holders, targets and polar expressions were not carried over. The seven researchers annotated in rounds of 50 comments each. In addition to this, annotators received small sets of 20 comments for quality control and Inter-Annotator Agreement (IAA) calculations. These datasets were annotated without discussions, but the results were discussed in order to resolve any ambiguity or potential issues.

4.1. Annotation Procedure

The sentiment annotation was performed at both the comment-level and the sentence-level. The annotators marked polarity (positive, negative) and intensity (slight, standard, strong). The original texts were sentence-segmented and tokenized using Stanza (Qi et al., 2020). Manual inspection of the resulting data shows that the tool provides accurate segmentation; however, cases such as emoticons (‘`:-D`’, etc.) are sometimes split erroneously. Sentences containing both positive and negative sentiment were annotated separately for both polarities. Sentences containing no polarity were marked as neutral.

Type	**Total**	**SMH**	**GP**
Comments	2 250	1 050	1 200
Sentences per comment	3.4	3.5	3.3
Words per sentence	14.8	16.0	13.8

Table 1: Number of comments, and average number of sentences per comment and words per sentence. We see that feedback in psychiatric care tends to be longer in both measures.

Comments were generally annotated using the same set of guidelines as for sentences, but comment-level polarity was scored based on a general impression of the comment as a whole, not just an aggregate of the sentiment of the sentences it contains. Annotators placed special weight on how the actual service (GP, SMH) was evaluated when assigning labels at the comment-level. Basic comment statistics are reported in Table 1.

Examples The comments in the dataset vary in terms of how sentiment is expressed, and sentiment expressions can contain direct evaluations, as we see in Example (1), where we find the strong positive adjective *fantastisk* ‘fantastic’ describing hospital employees, as well as the adjective *fin* ‘nice’ describing the patient’s stay.

(1) *Fantastiske ansatte som har gjort*
fantastic employees who have made
oppholdet så fint .
stay.the so nice .
‘Wonderful employees that have made my stay so nice.’

In Example (2) on the other hand, we see a more indirect evaluation of the treatment. The suggestion that a video consultation is useful is interpreted as a slightly negative evaluation of the health service in question.

(2) *Kanskje kunne videokonsultasjon være til*
Maybe could video.consultation be to
nytte ved en slik livssituasjon ?
use by a such life.situation ?
‘Perhaps a video consultation could be of use in this kind of life situation?’

The data also contain several neutral examples – typically patients reporting on their own health situation, as in Example (3), or general descriptions.

(3) *I tillegg bruker jeg øyedråper mot*
In addition use I eye.drops against
høyt trykk (grønn stær) .
high pressure (glaucoma) .
‘In addition, I use eye drops for high pressure (glaucoma).’

Further Discussions The annotators had weekly meetings both with each other, and with researchers working on sentiment analysis, to discuss problematic cases and to update the guidelines so that they better reflect the choices taken. Some of the issues discussed include to what extent descriptive sentences can indicate the patients' opinion, and to what extent context should be taken into account. In cases where sentiment would be ambiguous without context, annotators were asked to use the full context of the comment, and also their knowledge of the field. In some cases, apparently negative words could indicate positive sentiment, as in cases where patients note that they get diagnosed with a serious illness, perhaps indicating that it was good that it was actually diagnosed, or that it reflects well on the GP who diagnosed them correctly, and not focusing on the negative aspect of having a disease in itself.

4.2. Style and Variation

The data is notably different from the existing resources for Norwegian sentiment analysis. Being user-generated language, without editing, it contains some variation, both grammatical and stylistic. Certain terms pertaining to the domains, such as *fastlege* 'general practitioner', and *opphold* 'stay' are naturally very frequent. We also observe a tendency for certain polar expressions to be very frequent, with examples including *fornøyd* 'pleased' being among the top 20 most frequent words in the training set. A similar tendency is not observed in the NoReC$_{fine}$ data, where there is more variation and a weaker tendency for certain polar expressions to dominate. However, the texts notably do not contain much medical jargon, although some terms related to diseases and treatments. We believe the texts to be different from typical *medical domain* writing. The texts further reflect the society to some degree: both official Norwegian written norms are represented, Bokmål and Nynorsk. There is also some English in the dataset. Finally, some comments show indications of lexical and syntactic patterns associated with learner language, and with spoken language in general. For example, subjectless sentences are relatively common, as in Example (4), where the subject of the verbs *har* and *kan* has been elided.

(4) *Har vert innlagt før og kan trygt*
Have been admitted before and can safely
si at her burde noen gripe tak .
say that here should some grab hold .
'[I] have been admitted before and can safely say that someone should address the problems here.'

	POS	NEG	NEUT	MIX
Train	1 396	1 753	476	220
Test	1 396	1 755	477	220
Total	2 792	3 508	953	440

Table 2: Class distribution in the dataset for Positive, Negative, Neutral and Mixed sentences. The high similarity in counts is due to all classes being weighted when creating the splits, making sure both splits were as similar as possible.

4.3. Data Preparation and Splits

Although the full dataset is annotated with information about intensity, we decided not to include these attributes for the experiments in the current study due to the added complexity. Sentences containing slight, standard or strong polarity are thus labeled as only belonging to one of the two polarity classes, POS or NEG. Sentences containing both of these, in any polarity constellation, are labeled as MIX, while sentences with no polarity are labeled NEUT. The distribution of these four classes in the dataset is reported in Table 2. As we can see, the class distribution is well balanced across the training and test data, which were split 50–50, motivated by the need to ensure the test set is large enough and the fact that we do not perform any training. The split was done making sure that both datasets contained a similar number of labels for each class. The sentences were randomly sampled for each class. We mainly focus our experiments on this four-class setup, because we consider this to correspond most closely to a realistic use-case and because distinguishing sentiment-containing data from neutral sentences is an important part of naturally occurring data. However, we also report results for a reduced 2-class version, discarding the MIX and NEUT examples.

4.4. Human Inter-Annotator Agreement

First Round Initially, two test rounds of annotation were performed, in order to evaluate the guidelines, the experience of the annotator, and their agreement. Following an initial set of test annotations, smaller sets of 20 comments were annotated with regular intervals to judge the annotators' progress. Although the results from the first of these sets showed a large variance between annotators, with kappa agreement scores for annotations including intensity varying from very low at 0.21 to as high as 0.76. We note that much of the disagreements stem from the annotations of intensity. However, when considering to what extent annotators

	A1	A2	A3	A4	A5	A6
A1	1.0	0.96	0.95	0.95	0.96	0.95
A2	0.96	1.0	0.95	0.92	0.95	0.92
A3	0.95	0.95	1.0	0.95	0.95	0.93
A4	0.95	0.92	0.95	1.0	0.92	0.93
A5	0.96	0.95	0.95	0.92	1.0	0.92
A6	0.95	0.92	0.93	0.93	0.92	1.0

Table 3: Inter annotator agreement scores for positive, negative and neutral sentiment from the second round of IAA-annotations. One annotator who annotated for the project did not partake in these specific rounds.

agree on the two polarity classes and neutral sentences, ignoring intensity, the lowest kappa score was 0.62, and the highest 0.93.

Second Round Following the results from the first of the 20-comment annotations, the annotators discussed various specific cases of disagreement. Topics of discussion included to what extent certain expressions exhibit sentiment, and annotators were asked to give justification for their annotations, which were then discussed before consensus was made, if possible. After a series of discussions, a new IAA dataset was annotated to show to what extent these discussions affected the agreement among annotators, giving the results in Table 3, not including intensity, where we see that scores have improved markedly, showing that humans agree on which sentences are positive, negative and neutral. There is notably more disagreement for the negative polarity than for the positive. We also observe that there are often more than one annotator disagreeing for negative polarity, indicating greater variation and uncertainty for negative labels. One annotator annotated for the project, but did not partake in these two IAA-rounds.

5. LLM-Based Annotation

For free-text comment analysis, finding human annotators can be difficult. Not only are there few to choose from who have the necessary background knowledge to interpret the context correctly, annotation work can be expensive and time-consuming, placing additional economic strain on health services researchers or other personnel annotating data. We want to explore whether newer models could aid in this effort. Therefore, we compare the zero-shot and few-shot performance of two Norwegian LLMs with the annotations of six healthcare professionals.

5.1. Language Models

In particular, we evaluate predictions of one relatively small but instruction-finetuned model, ChatNorT5, and of a larger model that has not been finetuned on any downstream tasks, NorMistral.

ChatNorT5 This model is an instruction-finetuned version of `nort5-large`, an 808-million-parameter Norwegian encoder-decoder language model (Samuel et al., 2023).[2] By itself, NorT5 is pretrained on masked language modeling (Raffel et al., 2020), therefore, we further finetune it on instructions (via causal language modeling), to turn it into a generative model capable of predicting sentiment labels in zero-shot or few-shot settings. To make the evaluation more realistic, we train on a general set of instructions, not specifically focusing on sentiment analysis. We use a collection of Alpaca-like datasets (Taori et al., 2023) and translate them from English to Norwegian Bokmål with OPUS-MT (Tiedemann and Thottingal, 2020).[3] In total, we have translated 287k conversations and then finetuned the model for one epoch. One conversation consists of multiple query-response turns, and the model is trained to produce a gold response (using its decoder part) given all previous turns (provided to the encoder part).

NorMistral In addition to the instruction-tuned model, we also test an openly available Norwegian language model called `normistral-7b-warm`.[4] This model has been continually pretrained from the Mistral 7B model (Jiang et al., 2023) and has been shown to perform well in zero-shot and few-shot evaluations, even without instruction-tuning.

5.2. Likelihood Scores

As using the generated output from a causal LLM can lead to difficulties in mapping to the gold classes, we only consider a fixed set of possible responses for each prompt. We follow Brown et al. (2020) and formulate sentiment analysis as a task of choosing one prompt completion out of a limited number of other possible completions, based on

[2] https://huggingface.co/ltg/nort5-large

[3] UltraChat (Ding et al., 2023): https://huggingface.co/datasets/stingning/ultrachat, ShareGPT: https://huggingface.co/datasets/philschmid/sharegpt-raw, WizardLM: https://huggingface.co/datasets/WizardLM/WizardLM_evol_instruct_V2_196k and SODA (Kim et al., 2022): https://huggingface.co/datasets/allenai/soda.

[4] https://huggingface.co/norallm/normistral-7b-warm

their *likelihood scores*. Both evaluated language models can output $P(s_i \mid s_{0:i})$, the estimated probability of producing a subword s_i given the previous subwords $s_{0:i} = s_0 \ldots s_{i-1}$. We use this ability to test three ways of calculating a likelihood score of a completion $c = (c_0 \ldots c_n)$ given a query q:

1. $\mathcal{L}_1(c \mid q) = \sum_{c_i \in c} logP(c_i \mid q, c_{0:i})$,
2. $\mathcal{L}_2(c \mid q) = {}^1/_n \cdot \sum_{c_i \in c} logP(c_i \mid q, c_{0:i})$,
3. $\mathcal{L}_3(c \mid q) = {}^1/_{n_{\text{char}}} \cdot \sum_{c_i \in c} logP(c_i \mid q, c_{0:i})$.

The first formula calculates the actual estimated log-probability of c given q; however, in practice, this formulation tends to overestimate the likelihood of short sequences. Therefore, we try to normalize the likelihood by the length of completion c – the second formula normalizes by its number of subwords, n, and the third formula by its number of characters, n_{char}.

5.3. Prompting

As there is little earlier work on prompts for sentiment analysis for Norwegian, we based our initial prompts on the existing sentiment-related prompts in the FLAN dataset (Wei et al., 2022). The FLAN dataset contains four datasets: IMDB (Maas et al., 2011), Sent140 (Go et al., 2009), SST-2 Socher et al. (2013) and Yelp.[5] FLAN includes several sets of English SA prompts for each of these, and they were manually translated into Norwegian.

Prompt Variation The prompts from FLAN contain variation in terms of multiple choice variation, differences in formality, as well as different near synonyms, and the words used to refer to the text itself (*the preceding*, *this*). We aimed at keeping some variation, but discarded multiple choice questions and informal variants, and did not experiment with synonyms.

Prompt Filtering Among all Norwegian translations of prompts, only natural-sounding sentences were considered. As we calculate likelihoods based on a certain reply given a prompt, we also wanted to keep the number of possible replies low. Some sentences were discarded due to requiring very different replies. We also wanted to be able to compare the prompts with each other, and therefore excluded sentences that would force us to expand our number of prompts drastically. FLAN-prompts not related to sentiment classification were excluded. The resulting base prompts are shown in Table 4.

[5] https://course20.fast.ai/datasets.html

Prompt Expansion These base prompts were then expanded to create 48 prompts. Each resulting prompt is given a 4-number code based on the kind of modification it received. The first number (1-9) indicates the base prompt from Table 4. The second number indicates whether the test set sentence comes after (1) or before (2) the prompt. The third number indicates whether the question has no mentions of any of the four classes (0), mention of positive and negative (2), or all four (4). The final number indicates whether the word *positiv* comes before (2) or after (1) the word *negative*. We give an example of prompt `8-2-4-2` in Example (5). The id means that it is based on prompt 8, has the target sentence before the prompt, has 4 classes, and has *negativ* before *positiv*.

(5) *Oppfatter du denne setningen som*
Consider you this sentence.the as
positiv, negativ, blandet eller nøytral?"
positive, negative, mixed or neutral?
'Do you consider this sentence positive, negative, mixed or neutral?'

Possible Replies to Prompts Each prompt is combined with a limited set of possible answers. Much of the variation in these replies comes from the two main classes of answers, one with the word *setningen* 'the sentence' and one containing *sentimentet* 'the sentiment', which require masculine (*setningen*) and neuter (*sentimentet*) agreement, respectively. We introduce versions of replies that have a pronoun (*den*, 'it') instead, and finally versions where only the class is mentioned in the reply. In total, there are 16 alternatives per prompt, of which two have no difference depending on grammatical gender, giving 30 different replies. Each possible reply was associated with its suitable prompt data. A prompt file containing all possible prompts with all possible replies for each prompt was the basis of our experiments.

Examples6, 7 and 8 show how the answers to a prompt vary in syntactic structure and content. Example 6 shows a full sentence referring back to the sentence (or sentiment), while in 7 it is substituted by a pronoun, and finally in 8 both the pronoun and the verb is elided to provide a minimal answer.

(6) *Setningen er positiv .*
the.sentence is positive .
'The sentence is positive.'

(7) *Den er positiv .*
it is positive .
'It is positive.'

(8) *Positiv .*
positive .
'Positive.'

ID	Norwegian	English
1	Hvordan er sentimentet til denne setn.?	How is the sentiment of this sent.?
2	Hva er sentimentet til denne setn.?	What is the sentiment of this sent.?
3	Hvordan vil du beskrive sentimentet til denne setn.?	How would you describe the sentiment of this sent.?
4	Beskriv sentimentet i denne setn..	Describe the sentiment in this sent..
5	Ville du sagt at denne setn. er positiv eller negativ?	Would you say that this sent. is positive or negative?
6	Vil du si at denne setn. er positiv eller negativ?	Would you say that this sent. is positive or negative?
7	Er sentimentet i denne setn. positivt eller negativt?	Is the sentiment in this sent. positive or negative?
8	Oppfatter du denne setn. som positiv eller negativ?	Do you see this sent. as positive or negative?
9	Er denne setn. positiv eller negativ?	Is this sent. positive or negative?

Table 4: The 9 base prompts, and their English translations. Note that the translations here are back-translations. sent.=sentence, setn. =setningen.

6. LLM Experiments and Results

We performed experiments for both zero-shot and few-shot set-ups for both ChatNorT5 and NorMistral. We look at both a 4-class representation and a reduced binary representation, which corresponds roughly to cases where we would expect low and high agreement, respectively. To evaluate the binary dataset, we simply do not evaluate the model output on the neutral and mixed labels, and we limit the evaluation by only investigating the predicted likelihoods for the replies mapping to the classes pos and neg. The results are compared to a simple bag-of-words Naive Bayes model baseline.

	$\mathcal{L}$	ChatNorT5	NorMistral
4-class	$\mathcal{L}_1$	39.9	**9.4**
	$\mathcal{L}_2$	40.6	2.8
	$\mathcal{L}_3$	**42.4**	2.7
2-class	$\mathcal{L}_1$	88.7	84.8
	$\mathcal{L}_2$	89.2	89.0
	$\mathcal{L}_3$	**89.3**	**89.1**

Table 5: Zero-shot results. The highest F_1 scores (among different prompts) for the 4-class and 2-class evaluation using the ChatNorT5 and NorMistral models, and the three formulations of likelihood scores.

Naive Bayes Baseline We set up a simple Naive Bayes model using the Natural Language ToolKit (NLTK) Python library (Bird et al., 2009), using the entire vocabulary of the train set, removing the 20 most common words. With this baseline, we achieve a macro F_1 score of 41.0 for the four-class problem, higher than the random baseline of F_1 22.0 For the binary setup we achieve a quite high F_1 score of 79.0 compared to the baseline of 50.0

Experimental Setup Given the document with all 48 prompts and 16 alternatives per prompt, we estimated the likelihoods for each sentence in the test set. For a given sentence in the test set, there are 16×3 likelihood measures. For each of the three different likelihood scores, we selected the maximum across these 16, mapped the reply alternatives to one of the four classes, and treated that as the predicted value of that sentence.

In the 2-class setup, only responses that map to the binary classes were iterated through. We could then calculate the macro F_1 score across the dataset for each likelihood, for each prompt. We use the likelihood method that gives the overall best F_1 score for all prompts in the test set to evaluate which prompts we use for the few shot setup. We then select the best-performing prompts from each model.

6.1. Zero-Shot Runs

In the zero-shot setup, we provide the model with a prompt, and calculate the likelihood for the 16 possible answers, using the likelihood estimates described above. Results for zero-shot runs with the two models, for both the 4-class and 2-class results, are reported in Table 5. We observe notable differences both between the two models, and between the binary and 4-classes. The ChatNorT5 model performs much better in the 4-class setting, but this is mainly due to high scores for the negative and positive classes. We find that neutral and mixed are difficult for both models.

ChatNorT5 We see that for ChatNorT5, it is the third likelihood that gives the best results, both in the 4-class and 2-class setup. The differences between the three values are not large, and we observe that the difference between the various prompts are far more marked. The two best overall prompts were `7-2-2` and `7-1-2`, and the best 4-class prompts were `7-2-4-2` and `8-2-4`. `7-2-2` was also the best binary prompt, along with `5-1-2`. We see that three of these are based on prompt 7, which is originally binary but expanded to 4-class in

`7-2-4-2`. Regarding placement of the input sentence, in 19 of 24 pairs of sentences, having the sentence in front gives a higher macro F_1. In general, there is a problem that many prompts lead to few predictions of the neutral and mixed classes. We also see that, in terms of difference between the likelihoods, invariably, if a prompt predicts overall more positive sentiment for one likelihood, it does so for all likelihoods. It seems that prompts that work well for the positive and negative classes outperform other prompts even if they predict more neutral and mixed classes.

NorMistral For NorMistral we see that all prompts almost exclusively predict the mixed class. Here, the difference is larger between the three likelihood scores. While the likelihood score $\mathcal{L}_1$ gives the best 4-class score, the best binary score is obtained again by $\mathcal{L}_3$. While $\mathcal{L}_3$ has no predictions outside the mixed class, $\mathcal{L}_2$ has a single prediction outside, but $\mathcal{L}_1$ has 201 positive, 416 negative, and 5612 neutral predictions. The weak 4-class results are somewhat surprising, and can indicate either that the current prompts are largely inadequate for this model or that the model does not understand more nuanced sentiment. However, in the binary setup, the opposite is true. We approach very high F_1 scores, suggesting that weak 4-class scores are a result of inadequate prompting. Interestingly, the best-performing sentences are four-class prompts: `9-1-4` and `6-1-4`.

Due to low and largely similar results for the 4-class setup, we select only four prompts for NorMistral, exclusively from the binary setup, and also include prompts `2-1-0` and `6-1-2-2`.

6.2. Few-Shot Runs

Having run zero-shot runs for both models, we use the best performing prompts in a four-shot setting. This setup consists of four pairs of query–response examples, one from each class, given to the model as context, before the test sentence we want to make a prediction for. Each example was randomly sampled from the training set, and selected separately for each sentence in the dataset. These examples are all taken from the training set. The best scores for the three likelihoods are reported in Table 6.

Few-Shot with ChatNorT5 The results from the few-shot runs with ChatNorT5 are reported in Table 7. We observe that the F_1 score drops due to the 7-based prompts almost exclusively predicting the mixed class, while the 8-based prompt favors the neutral class, but is more balanced. In the binary setup, however, we see very high scores, almost beating the zero-shot results.

	$\mathcal{L}$	ChatNorT5	NorMistral
4-class	$\mathcal{L}_1$	28.6	**12.2**
	$\mathcal{L}_2$	28.6	2.7
	$\mathcal{L}_3$	28.6	2.7
2-class	$\mathcal{L}_1$	89.1	**84.9**
	$\mathcal{L}_2$	**89.3**	83.9
	$\mathcal{L}_3$	83.6	83.9

Table 6: Few-shot results. The highest F_1 scores for the 4-class and 2-class, as for the zero-shot results.

Prompt ID	4-class	2-class
`7-2-2`	3.8 (0.2)	89.0 (0.3)
`7-1-2`	3.2 (0.1)	85.2 (0.3)
`7-2-4-2`	7.2 (0.2)	87.8 (0.4)
`8-2-4`	26.1 (0.3)	83.0 (0.3)
`5-1-2`	28.3 (0.2)	62.0 (0.4)

Table 7: Mean F_1 and standard deviation for the few-shot experiments using ChatNorT5.

Few-Shot with NorMistral The results of the four prompts from the NorMistral zero-shot run are reported in Table 8. We see that NorMistral also struggles with making reliable predictions in the 4-class setup, but performs well on pos and neg, albeit not as well as ChatNorT5.

Comparison with Baseline We observe that for the 4-class setup, ChatNorT5 achieves similar scores to the baseline model in the zero-shot runs, while NorMistral achieves notably lower scores for all runs using the 4-class setup. However, for binary sentiment, both generative models achieve higher scores than the baseline in most cases, both for zero-shot and few-shot. The low scores in the 4-class setup are surprising for both models, and we hope to investigate this in later experiments.

Prompt ID	4-class	2-class
`2-1-0`	6.5 (0.4)	84.4 (0.1)
`6-1-4`	11.8 (0.2)	76.4 (0.5)
`6-1-2-2`	11.8 (0.2)	72.3 (0.3)
`9-1-4`	11.0 (0.1)	69.5 (0.5)

Table 8: Mean F_1 and standard deviations for the few-shot runs using NorMistral.

Prompt ID	A1	A2	A4	A6	A7
`2-1-0` (2)	0.69	0.00	0.70	0.62	0.70
`7-2-2` (4)	0.46	0.32	0.52	0.49	0.55
`7-2-2` (2)	0.79	0.00	0.81	0.76	0.83
`9-1-4` (2)	0.81	0.00	0.77	0.75	0.83

Table 9: Annotation agreement between the best predictions and the human annotators. Annotators A3 and A5 are not represented due to lack of applicable data in the test set. Prompt 7-2-2 is tested both in the 2-class and the 4-class setup. Annotator A7 did not partake in the previous IAA rounds.

6.3. Comparison with Human Annotators

Due to the inability of any prompt to reliably predict the NEUT and MIX classes, we selected prompts based on the best 2-class results, along with the best four-class and binary for zero shot CHATNORT5. This gives us three prompts: `7-2-2`, `2-1-0` and `9-1-4`.

Model-annotator IAA We compared these three combinations of prompts with models, comparing them individually with each annotator, which have the results presented in Table 9. We treat all documents annotated by an annotator as representing that annotator, and calculate kappa scores between the model and the human annotators, like we did with the human annotators. For the full 4-class problem, we see that agreement is low, but still not as low as some of the project-initial disagreements.

Human versus Model While we found the F_1 scores to much higher for the 2-class setup, treating the model outputs as annotations still gives us lower scores than that expected from humans familiar with the task. The obvious issue is that we struggle to get our models to reliably distinguish between neutral and polar sentences. Inspection shows that a prevalent error for both prompt `2-1-0` and `7-2-2` is to mistake positive sentences for negative, while prompt `9-1-4` has more cases where the model treats positive as negative. The most common mistake in the 4-class setup for `7-2-2` is that non-mixed sentences are classified as mixed.

7. Conclusion

This paper has described how free-text comments in patient surveys collected by the Norwegian Institute of Public Health have been annotated with information about sentiment. Specifically, our data comprise patient comments in Norwegian on experiences with General Practitioners and Special Mental Healthcare, which we have annotated with positive/negative polarity (including intensity) on both the comment- and sentence-level. In addition to describing the annotation guidelines and presenting an analysis of the resulting dataset, we also include experimental results on augmenting the human annotations with predictions by two different open-source pretrained large language models (LLMs); CHATNORT5 and NORMISTRAL. We report results for both zero- and few-shot settings for several different prompting configurations. We find that the predictions of the LLMs are sensitive to the particular prompt used, and that the best configuration depends on the specific model. Moreover, while we find that both models perform well for the simple binary cases where sentences are either positive or negative, they both struggle with neutral and mixed-polarity examples. Our error analysis shows that the predictions of the LMMs used in this study are still inferior in quality to the human annotations for our dataset.

8. Limitations

Annotator representations Due to work load limitations, we were not able to provide an even distribution of data across human annotators. This makes the claims on some of the annotators hold less than for others.

Intensity While we would have liked to include intensity scores, this will have to be the subject for later research. While interesting due to being a source of disagreement in humans, and we believe that differences in the treatment of intensity might reveal further differences between humans and models, it requires more space than what we could dedicate in this paper.

Variation We note that there is linguistic variation in the dataset, but addressing this is outside the scope of our paper. We hope to be able to return to this to be able to better assess how user language might affect how patients' voices are analyzed using systems often trained and evaluated on normative and edited language.

9. Ethical Considerations

While it might be possible to get similar or even better results with certain commercial models, there are several reasons why we opt for open-source Norwegian models. First of all, these models can be run locally, also without using APIs that would require sending data to servers not cleared for storage of our data, and do not pose any conflict in terms of privacy or potential data leakage. Secondly, models trained on data from the same area as the patients might lead to less likelihood of cultural bias affecting judgements. Finally, the models' training data are open, and it is therefore possible to investigate biases and potential problems should they arise.

10. Acknowledgements

We would like to thank Karoline Aasgaard, Kirsten Danielsen, Lina Harvold Ellingsen-Dalskau, Hilde Hestad Iversen and Inger Opedal Paulsrud for their annotation work. This work was supported by the Norwegian Research Council, through the project *Styrking av pasientstemmen i helsetjenesteevaluering: maskinlæring på fritekstkommentarer fra spørreundersøkelser og nettkilder.* 'Strengthening the patient voice in health service evaluation: machine learning on free-text comments from surveys and online sources', project number 103707. 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.

11. Bibliographical References

Farrokh Alemi, Manabu Torii, Laura Clementz, and David C. Aron. 2012. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Quality Management in Healthcare*, 21(1).

George Alexander, Mohammed Bahja, and Gibran Farook Butt. 2022. Automating large-scale health care service feedback analysis: Sentiment analysis and topic modeling study. *JMIR Med Inform*, 10(4):e29385.

Rebecca Anhang Price, Marc N Elliott, Alan M Zaslavsky, Ron D Hays, William G Lehrman, Lise Rybowski, Susan Edgman-Levitan, and Paul D Cleary. 2014. Examining the role of patient experience surveys in measuring health care quality. *Med Care Res Rev*, 71(5):522–554.

Mohammed Bahja and Mark Lycett. 2016. Identifying patient experience from online resources via sentiment analysis and topic modelling. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, Shanghai, China.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Kristina Doing-Harris, Danielle L Mowery, Chrissy Daniels, Wendy W Chapman, and Mike Conway. 2017. Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annu Symp Proc*, 2016:524–533.

Cathal Doyle, Laura Lennox, and Derek Bell. 2013. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*, 3(1).

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, CS224N project report, Stanford.

Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res*, 15(11):e239.

Rachel Grob, Mark Schlesinger, Lacey Barre, Naomi Bardach, Tara Lagu, Dale Shaller, Andrew Parker, Steven Martino, Melissa Finucane, Jennifer Cerully, and Alina Palimaru. 2019a. What words convey: The potential for patient narratives to inform quality improvement. *The Milbank Quarterly*, 97:176–227.

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. 2019b. What words convey: The potential for patient narratives to inform quality improvement. *The Milbank Quarterly*, 97(1):176–227.

Jared B Hawkins, John S Brownstein, Gaurav Tuli, Tessa Runels, Katherine Broecker, Elaine O Nsoesie, David J McIver, Ronen Rozenblum, Adam Wright, Florence T Bourgeois, and Felix Greaves. 2016. Measuring patient-perceived quality of care in us hospitals using twitter. *BMJ Quality & Safety*, 25(6):404–413.

John W. Huppertz and Peter Otto. 2018. Predicting hcahps scores from hospitals' social media pages: A sentiment analysis. *Health Care Management Review*, 43(4).

Hilde Iversen, Øyvind Bjertnæs, and Kjersti Skudal. 2014. Patient evaluation of hospital outcomes: An analysis of open-ended comments from extreme clusters in a national survey. *BMJ open*, 4:e004848.

Hilde Hestad Iversen, Mona Haugum, and Øyvind Bjertnæs. 2022. Reliability and validity of the psychiatric inpatient patient experience questionnaire – continuous electronic measurement (PIPEQ-CEM). *BMC Health Services Research*, 22(1):897.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Salud M. Jiménez Zafra, María Teresa Martín Valdivia, Isa Maks, and Rubén Izquierdo Beviá. 2017. Analysis of patient satisfaction in Dutch and Spanish online reviews.

Mustafa Khanbhai, Patrick Anyadi, Joshua Symons, Kelsey Flott, Ara Darzi, and Erik Mayer. 2021. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health & Care Informatics*, 28(1).

Mustafa Khanbhai, Leigh Warren, Joshua Symons, Kelsey Flott, Stephanie Harrison-White, Dave Manton, Ara Darzi, and Erik Mayer. 2022. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. *International Journal of Medical Informatics*, 157:104642.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.

Marte Kjøllesdal, Thor Indseth, Hilde Hestad Iversen, and Øyvind Bjertnæs. 2020. Patient experiences with general practice in norway: a comparison of immigrant groups and the majority population following a national survey. *BMC Health Serv Res*, 20(1):1106.

Larisa Kolesnichenko, Erik Velldal, and Lilja Øvrelid. 2023. Word substitution with masked language models as data augmentation for sentiment analysis. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 42–47, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Petter Mæhlum, Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 121–130, Turku, Finland. Linköping University Electronic Press.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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 hospital perspective. *International Journal 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*, (7.23).

Daniela Rohde, Mona Isazad Mashinchi, Nina Rizun, Dritjon Gruda, Conor Foley, Rachel Flynn, and Adegboyega Ojo. 2022. Generating actionable insights from free-text care experience survey data using qualitative and computational text analysis: A study protocol. *HRB Open Research*.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Marieke M. van Buchem, Olaf M. Neve, Ilse M. J. Kant, Ewout W. Steyerberg, Hileen Boosman, and Erik F. Hensen. 2022. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (aiprem). *BMC Medical Informatics and Decision Making*, 22(1):183.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian review corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Richard Wagland, Alejandra Recio-Saucedo, Michael Simon, Michael Bracher, Katherine Hunt, Claire Foster, Amy Downing, Adam Glaser, and Jessica Corner. 2016. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. *BMJ Quality & Safety*, 25(8):604–614.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations*, online.