Mixed Feelings: Cross-Domain Sentiment Classification of Patient Feedback

Egil Rønningstad*, **Lilja Charlotte Storset***, **Petter Mæhlum, Lilja Øvrelid, Erik Velldal**
Department of Informatics, University of Oslo, Norway
{egilron, liljacs, pettemae, liljao, erikve}@uio.no

Abstract

Sentiment analysis of patient feedback from the public health domain can aid decision makers in evaluating the provided services. The current paper focuses on free-text comments in patient surveys about general practitioners and psychiatric healthcare, annotated with four sentence-level polarity classes – positive, negative, mixed and neutral – while also attempting to alleviate data scarcity by leveraging general-domain sources in the form of reviews. For several different architectures, we compare in-domain and out-of-domain effects, as well as the effects of training joint multi-domain models.

1 Introduction

Sentiment analysis (SA), the computational analysis of opinions and emotions expressed in text, is one of the applications of natural language processing (NLP) that have found the most widespread use across many different areas, including medical domains (Yadav et al., 2018). As the task is mostly approached as one of supervised learning, access to sufficient amounts of labeled data is the main driver of performance. However, as manual annotation is costly, labeled data also represents a main bottleneck. For this reason it is typically desirable to be able to reuse existing resources when developing SA tools for a new area of application. Unfortunately, domain-sensitivity is a well-known effect across many different NLP tasks. Models trained on data from one domain (or genre or text-type) often underperform when applied to another due to variations in language use, terminology, and contextual nuances (Al-Moslmi et al., 2017; Gräßer et al., 2018).

*The authors contributed equally.

This paper investigates cross-domain effects in polarity classification of public health data, more specifically free-text comments from patient surveys for general practitioners and psychiatric healthcare providers. We here investigate the usefulness of data from a different domain and genre, i.e. professionally authored reviews collected from Norwegian news publishers. The datasets are annotated at the sentence level with the same four-class polarity labels; positive, negative, mixed, and neutral. In the following, we compare non-neural and neural architectures in both in-domain and cross-domain settings with the goal of providing high-quality sentiment analysis for Norwegian patient comments.

2 Datasets

We here briefly describe the two annotated SA datasets that form the basis of our experiments, also discussing some of their key differences.

NorPaC For the health domain we will be using a dataset introduced by Mæhlum et al. (2024), comprising free-text comments from surveys conducted by the Norwegian Institute of Public Health (NIPH), as part of their so-called patient-reported experience measures (PREMs). The dataset is dubbed NorPaC – short for Norwegian Patient Comment corpus – and comprises two related subdomains, corresponding to feedback on General Practitioners (GPs) and Special Mental Healthcare (SMH), with a total of 7693 sentences (4002 from GP and 3691 from SMH) annotated for polarity.

The NorPaC dataset is a valuable accession to Norwegian corpora, as it gives valuable insights to the national public health system. The texts are written by patients after encounters with the healthcare system, and gives rise to language with an everyday character, such as sentences with a conversational tone or even incomplete sentences and spelling mistakes. Example 1 shows a positive

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

patient feedback sentence that is written solely in capital letters, in addition to containing a typing mistake in the personal pronoun *jeg*, 'I'. Example 2 shows a negative review with a colloquial tone, containing three exclamation marks at the end of the utterance.

(1) *FIKK HENVISNING DA JGE BA*
Got referral when (I) asked
OM DET, OG GÅR STADIG TIL
about it, and goes constant to
UTREDNING DER.
examination there.
'Got a referral when I asked for it, and am constantly going for examination there.'

(2) *Det er for dårlig!!!*
It is too bad!!!
'It is too bad!!!'

NoReC The Norwegian Review Corpus (NoReC; Velldal et al., 2018) comprises full-text reviews collected from major Norwegian news sources, covering a range of different domains (movies, music, literature, restaurants, various consumer products, etc.). We here use a version dubbed NoReC$_{fine}$ (Øvrelid et al., 2020), a subset of roughly 11,000 sentences across more than 400 reviews with fine-grained sentiment annotations, here aggregated to the sentence-level (Kutuzov et al., 2021) using the above-mentioned label set of four classes.[1] In contrast to NorPaC, the reviews are written by professional authors, meaning more creative writing but with sentences that are typically complete and grammatically correct.

(3) *Den er en pølse i salatens tid,*
It is a sausage in salad's.the time,
en slags mumlemanisk
a kind.of mumblemaninc
manns-modernitets-manifestasjon
man-modernity-manifestation
'It is a sausage in the age of the salad, a kind of mumble-manic male-modernity-manifestation'

Example 3 shows one of many creative sentences in NoReC. *En pølse i salatens tid*, 'a sausage in the age of the salad', is a figurative way to emphasize the fact that this movie is not among the trendy, i.e. 'the salad', but rather acts like 'a sausage'. Further, the author describes the movie as a *mumlemanisk manns-modernitets-manifestasjon*, 'mumble-manic male-modernity-manifestation'. This exemplifies the complexity of many of the texts in the NoReC dataset where authors may construct new and creative expressions.

Genre and text type The two datasets can be said to be found at opposite ends in terms of language and writing style. In contrast to the professionally authored reviews in NoReC, containing grammatically correct texts with higher complexity and creativity, the NorPaC patient comments consist of more colloquial language. It also comes with many of the other hallmarks of user-generated content, such as more frequent spelling mistakes and incomplete sentences, as well as unorthodox use of case and punctuation. While such properties will generally contribute to increasing the vocabulary size, NoReC still contains almost three times as many unique lemmas as NorPaC, due to the fact that it contains more creative and varied language (with a higher degree of figurative expressions, etc.), as mentioned above, in addition to covering multiple domains.

Class distribution Table 1 summarizes some relevant statistics for the two corpora, showing the number of examples across the four classes, as well as average token length of sentences.

For the NoReC reviews, we see that we have many more examples for the positive than the negative category. For the NorPaC patient feedback, in contrast, the negative category is notably larger, although the number of positive and negative examples are more balanced than in NoReC.

Another striking difference is the much higher ratio of neutral sentences in NoReC compared to NorPaC; 47% vs. 12%, respectively. This is not surprising if we consider the genre differences; professional reviews need to provide a lot of non-sentiment bearing background and descriptions of the object under review. The ratio of sentences with mixed polarity, however, is similar across the datasets, and is also the smallest sentiment class.

Related to the class distribution, we also observe some interesting differences with respect to the average token length of sentences. While the length is the same across the positive and negative sentences in the NoReC reviews, the length of negative sentences in the NorPaC patient comments tend to be substantially longer than the positive ones. However, for both datasets we see that neu-

[1] https://huggingface.co/datasets/ltg/norec_sentence

		Positive	Negative	Neutral	Mixed	Total
GP	Sentences	1265 (32%)	1903 (48%)	654 (16%)	174 (4%)	4002
	Avg. tokens	11.8	15.61	10.38	19.99	13.81
SMH	Sentences	1524 (41%)	1604 (44%)	291 (8%)	266 (7%)	3691
	Avg. tokens	13.1	18.48	10.53	23.68	15.94
NorPaC (GP+SMH)	Sentences	2789 (36%)	3507 (46%)	945 (12%)	440 (6%)	7693
	Avg. tokens	12.53	17.03	10.78	22.29	14.93
NoReC	Sentences	3514 (31%)	1663 (15%)	5393 (47%)	867 (8%)	11437
	Avg. tokens	18.57	18.18	13.78	25.92	16.78

Table 1: For each polarity class we show the distribution of number of sentences and average sentence length across the GP and SMH datasets within NorPaC, and for the NoReC dataset.

tral sentences tend to be shorter, while the mixed class displays substantially longer average length, which intuitively makes sense given that they per definition must express at least two opposing sentiments.

3 Experimental results

Below we report experimental results for a range of different models and architectures on the datasets described above. We start by providing details about the models and the experimental set-up, before discussing the results for both in-domain and cross-domain classification.

3.1 Models and experimental set-up

The NorBERT3 series of models (Samuel et al., 2023; Kutuzov et al., 2021) represent the 3rd generation of pre-trained Norwegian masked language models (MLMs) based on the BERT transformer architecture (Devlin et al., 2019). We fine-tuned text classifiers for two different sizes of NorBERT3 – Base and Large – with 123M and 353M parameters, respectively. GPU memory requirements were 8 and 35 GB. The NorT5 (Samuel et al., 2023) models are pretrained on the same Norwegian data as NorBERT3, and we fine-tune NorT5 Large to generate sentiment labels as a sequence-to-sequence task. NorT5 Large has 808M parameters. During fine-tuning with a batch size of 24, 71GB GPU memory was used. For all these models we report the mean weighted average F_1 over 3 runs. More details of the hyperparameter search are found in Appendix A. As a baseline, we also train a Support Vector Machine (SVM) model with a linear kernel and bag-of-words features.[2] The random baseline for the task yields an F1-score of between 22% and 23% for all training datasets, averaged across 1000 runs.

For NoReC we use the predefined data split, with 80-10-10 percentages respectively for the training, validation and test set. We define a similar split for NorPaC, randomly selected on the comment-level to make sure sentences from the same comment are not separated across splits, while also ensuring a balanced class distribution.

3.2 In-domain patient comment results

Table 2 shows results when training and testing on sentences from the NorPaC corpus. While the main focus of this section is to assess the in-domain performance of models trained on the NorPaC patient comments, recall that this corpus comprises two different sources; feedback regarding General Practitioners (GPs) and Special Mental Healthcare (SMH). We therefore also report results for training and testing on data from the individual sources separately – including cross-source training and testing.

We see that training on GP yields very strong test results: Not only are in-domain results for training and testing for SMH lower, but test results on SMH are competitive when training on GP. In the same vein, we see that for most models, joint training on the entire NorPaC data boosts results for SMH, with the only exception being NorBERT3 Large, where the best results for SMH are actually found when training on GP only (although the differences are marginal). In sum, we find that, within the NorPaC domain(s), the generalization capabilities of the GP-trained models

[2]The features correspond to the full vocabulary of the tokenized texts for each corpora, as preliminary experiments showed that best results were obtained without any feature selection or weighting (i.e. no TF-IDF, frequency cutoffs, etc.). The number of features range from approximately 5K for the GP/SMH models, through 8K for the full NorPaC data and 22K for NoReC, and finally 27K for NorPaC+NoReC.

Model	Train	Test		
		GP	SMH	NorPaC
SVM (BoW)	GP	**63.65**	66.42	64.96
	SMH	57.86	66.77	62.26
	NorPaC	62.90	**68.34**	**65.52**
NorBERT3 (Base)	GP	**84.13**	82.02	83.14
	SMH	79.43	82.96	81.22
	NorPaC	83.61	**83.34**	**83.49**
NorBERT3 (Large)	GP	85.79	**84.85**	**85.41**
	SMH	81.23	84.61	82.95
	NorPaC	**86.00**	84.28	85.22
NorT5 (Large)	GP	84.34	83.65	84.08
	SMH	81.03	84.24	82.70
	NorPaC	**85.03**	**85.05**	**84.54**

Table 2: Results for training and testing on the GP and SMH datasets within NorPaC.

Model	Train	Test	
		NorPaC	NoReC
SVM (BoW)	NorPaC	65.52	37.84
	NoReC	42.11	**54.42**
	NorPaC+NoReC	**66.20**	53.35
NorBERT3 (Base)	NorPaC	83.49	59.09
	NoReC	68.03	75.63
	NorPaC+NoReC	**83.71**	**76.14**
NorBERT3 (Large)	NorPaC	**85.22**	59.19
	NoReC	66.38	**78.88**
	NorPaC+NoReC	85.03	78.40
NorT5 (Large)	NorPaC	84.54	58.14
	NoReC	70.88	**76.73**
	NorPaC+NoReC	**85.06**	75.79

Table 3: Results for training and testing on sentences from both the NorPaC patient comments and the NoReC reviews.

are so good that the benefit of joint training on GP and SMH are less than anticipated. One contributing factor here might be that the GP data overall is written in a more explicit and straightforward manner compared to the SMH data, which might contain parts that are perceived as noisy for the model. Hence, training on GP and testing on PHV yields better results than vice versa. Finally, and as expected, we see that the neural models outperform the SVM model and that larger models generally tend to outperform smaller ones, although NorT5 Large actually tends to be outperformed by the smaller NorBERT3 Large model.

3.3 Cross-domain results

Table 3 shows results for several combinations of training and testing on both NorPaC and NoReC. First, we note that the in-domain results for NorPaC are substantially higher than the in-domain results for NoReC. This makes sense, given that NoReC in practice covers many different domains and has a much more diverse vocabulary than NorPaC. This observation most likely also has bearings on the cross-domain results, where we see a smaller relative drop in performance when testing the NoReC-trained models on NorPaC, than vice versa. Another contributing factor to the (expected) drops in performance for the cross-domain results can be the differences in the class distribution for the two datasets, as discussed above.

Turning to the joint training on the combination of NoReC and NorPaC, we again see that the test scores are substantially higher on NorPaC than NoReC for all models. For the NorBERT3 Base model, the joint training improves results across both datasets. However, for NorBERT3 Large, we see that the in-domain variants gives the highest scores for both datasets, but only by a small margin. For the SVM model, we see the same tendency with in-domain training on NoReC, yielding slightly better performance than joint training.

In an error analysis of in-domain vs. out-of-domain results for NorBERT3 Large evaluated on the NorPaC test set, we observe that the model trained on NorPaC is better at predicting negative sentences, compared to the model trained on NoReC. Here, the in-domain model classifies 92% of the negative samples correctly, whereas the out-of-domain model only identifies 39% of them. Out of the true negative samples, the NoReC-trained model predicts 59% of them as neutral. We believe the prediction of the negative class is the largest contributor to the lower performance of the NoReC-trained model, as this class makes up 46% of the NorPaC test set. However, there is one class for which this model performs slightly better than the in-domain model. As we recall from Table 1, the neutral class is the largest class in the NoReC dataset. This is most likely the reason why the NoReC-trained model classifies 95% of these instances in the NorPaC test set correctly, as opposed to the NorPaC trained model, which correctly classifies 69% of them. In sum, a closer

look at the per-class results reveals clear effects of the class distribution in the training set on model performance.

Learning curves for in-domain data To gauge the effect of the number of in-domain training examples, we computed learning curves where models are trained on partitions that are created by successively halving the NorPaC training set, with and without including the full NoReC training data. Figure 1 plots the effect on fine-tuning NorBERT3 Large. Utilizing only 6.25% (386 samples) of the NorPaC training set we find a strong performance gain of adding the cross-domain NoReC dataset. The effect is reduced, but present up to 50% (3087 samples). However, with the full NorPaC training set containing 6175 samples, we find that adding cross-domain data is harmful for the model performance. This shows how cross-domain data can help when in-domain datasets are small, but should not be added indiscriminately.

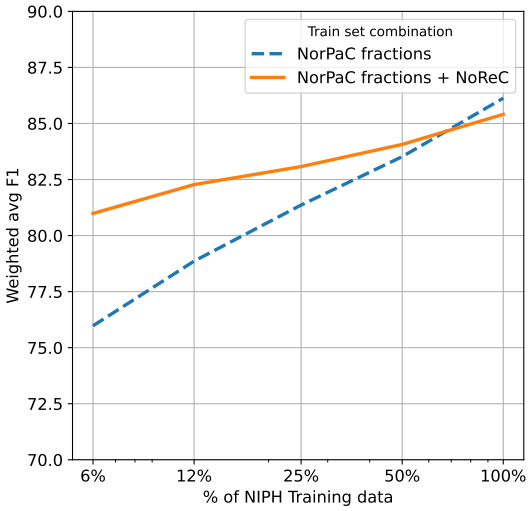

Figure 1: Learning curves, for two configurations: **NorPaC fractions**: The model is trained on fractions of the NorPaC training split, from 6.25% ≈ 6% (386 samples) successively doubling the training set up to the full NorPaC training split. **NorPaC fractions + NoReC**: The same fractions of the NorPaC training split, mixed with the full NoReC training split. All evaluations are on the full NorPaC test set, averaged over three runs with different seeds, and with the amounts of in-domain training data shown on log-scale.

4 Summary

This paper has reported experimental results for polarity classification of sentences in a Norwegian dataset dubbed NorPaC, comprising free-text comments from patient surveys collected as part of evaluating public healthcare services. In addition to assessing cross-domain effects between two healthcare sub-domains – feedback on general practitioners and psychiatric healthcare – we have also assessed the effect of leveraging general-domain sentiment annotations, based on the NoReC review data. Rather than just annotating the simple binary classification of positive/negative sentences, our datasets additionally indicate both neutral and mixed sentences. We show how several of our tested model configurations surpass 85% weighted F_1 for this four-class set-up. We also show how including out-of-domain data improves model performance when in-domain data is limited, but that better performance can be achieved with in-domain data alone once the the amount of annotated data crosses a critical threshold. Our analyses give new insights into both the NorPaC and NoReC datasets, including the differences and similarities between them.

Acknowledgments

This work was supported by two research projects funded by the Research Council of Norway (RCN), namely 'Sentiment Analysis for Norwegian Text' (SANT), funded by an IKTPLUSS grant from RCN (project no. 270908), and '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.

References

Tareq Al-Moslmi, Nazlia Omar, Salwani Abdullah, and Mohammed Albared. 2017. Approaches to cross-domain sentiment analysis: A systematic literature review. *IEEE Access*, 5:16173–16192.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *ACM International Conference Proceeding Series*, volume 2018-, pages 121–125, New York, NY, USA. ACM.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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.

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.

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.

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).

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018. Medical sentiment analysis using social media: Towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Hyperparameter tuning for NorBERT3-based models

We chose NorBERT3 base and large as the models to fine-tune for the text classification task. This model series has proven to perform well on previous comparisons for sentiment analysis on Norwegian sentences (Samuel et al., 2023). In order to find the best hyperparameters for our task, we first experimentally determine the best combination of learning rate and batch size. Table 4 shows the results for the two model sizes. All experiments are evaluated by accuracy on the development split, using the best of 10 epochs and one seed per hyperparameter combination. With the best performing settings for learning rate and batch size, we further search for improved performance by adjusting dropout in the classifier head, warm-up ratio and weight decay during fine-tuning. The search space for these hyperparameters are shown in Table 5. The best performing settings are shown in Table 6. The final choice of hyperparameters are shown in Table 7.

Model	lr	16	32	64
base	1e-05	**78.28**	77.82	77.76
base	2e-05	77.92	78.12	77.79
base	5e-05	76.09	77.04	78.18
large	1e-05	80.44	81.12	80.37
large	2e-05	80.96	**81.35**	80.89
large	5e-05	79.23	80.44	80.60

Table 4: Learning rate and batch size hyperparameter search for NorBERT3-base and large.

Model	Search space
classifier dropout	[0.05, 0.1, 0.25, 0.4]
warm-up ratio	[0.01, 0.05, 0.1, 0.2]
weight decay	[0.001, 0.01, 0.1]

Table 5: Search space for classifier dropout, warmup ratio and weight decay for NorBERT3 base and large, after best learning rate and batch size was identified.

Model	Dropout	Wu_ratio	W_decay	Dev acc.
base	0.25	0.20	0.010	**78.77%**
base	0.10	0.20	0.010	78.71%
base	0.25	0.20	0.100	78.58%
large	0.25	0.10	0.100	**82.10%**
large	0.25	0.10	0.001	81.91%
large	0.40	0.20	0.001	81.84%

Table 6: Top-3 performing models, for NorBERT3 base and large, when searching for optimal parameters for classifier dropout, warm-up ratio and weight decay.

Model	Base	Large
batch size	16	32
learning rate	1e-05	2e-05
classifier dropout	0.25	0.25
warmup ratio	0.20	0.10
weight decay	0.01	0.10

Table 7: Final hyperparameters selected for the NorBERT3 base and large finetuning, as informed by our hyperparameter search. Other hyperparameters are left as their defaults.