

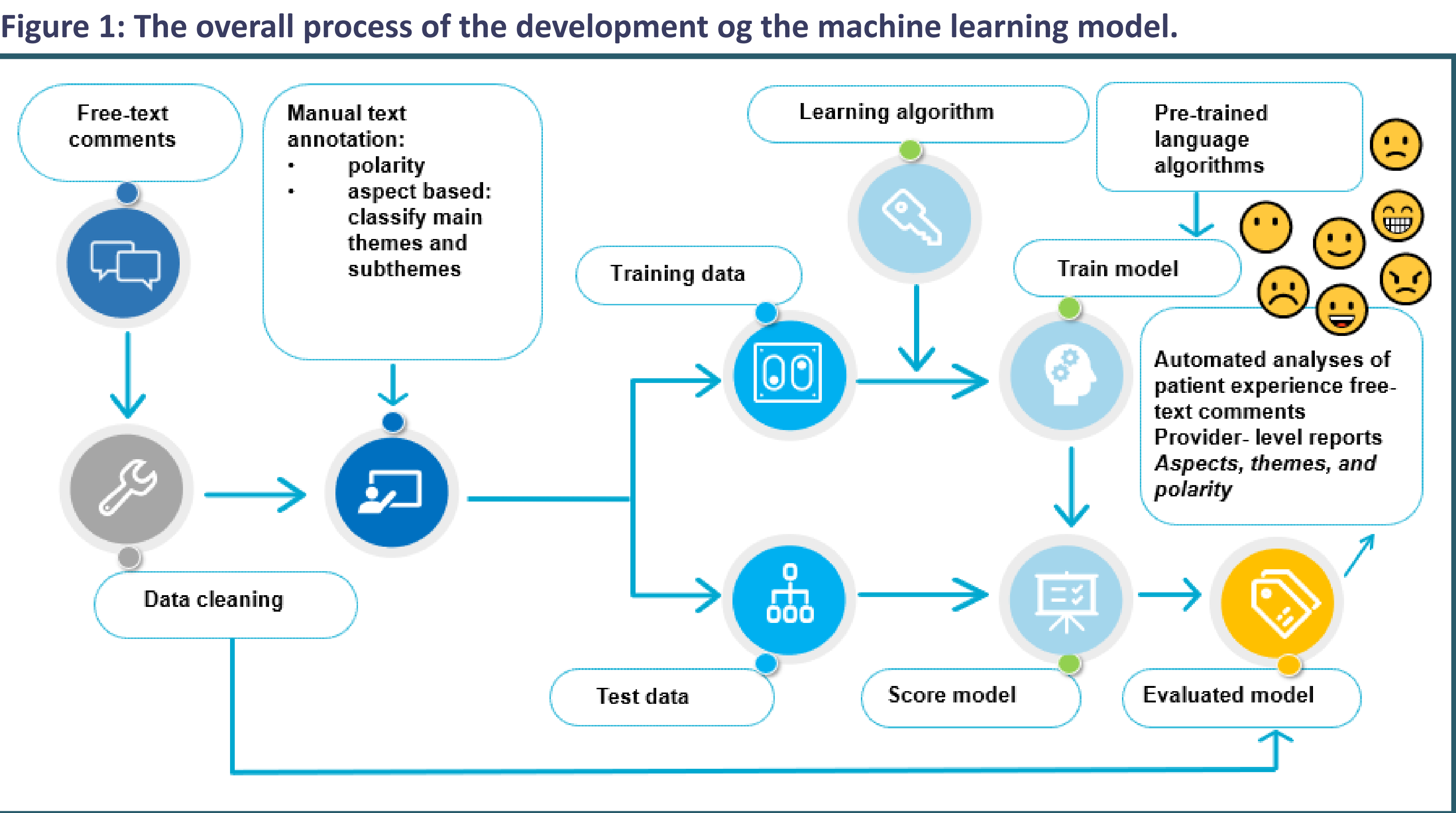
R. M. Norman¹, E. Jelin¹, P. Mæhlum², E. Velldal², L. Øvrelid², Ø. A. Bjertnæs¹

1: Norwegian Institute of Public Health, Division of health services, Cluster for health services research, Norway

2: University of Oslo, Department of Informatics, Section for Machine Learning, Language Technology Group, Norway

Introduction

The overall aim of the study is to develop and evaluate machine learning tools, using Natural Language Processing (NLP) techniques, including sentiment analysis, for automatic analysis of free-text patient experience feedback, and integrating the results in a new generation of provider level feedback. As the patient feedback data are different in terms of both domain and genre, they pose a challenge to existing models, but also allow us to evaluate challenges with patient feedback compared to more traditional datasets. A critical step is manual annotation of comments, which ensures the model's accuracy and effectiveness.



Aim

The aim of this part of the study is to detail the polarity annotation process of the comments, which serves as the foundation for developing and evaluating NLP-based tools for automatic sentiment analysis.

Methods

Free-text comments from two national patient experience surveys (general practice and mental health care) were independently annotated by three to six researchers in a structured two-phase process including marking comments for polarity and intensity.

The process includes several rounds of inter-annotator agreement (IAA) testing, employing Kappa statistics to ensure consistency and reliability. The annotation scheme is based on earlier research on sentiment analysis in Norwegian. As base models, both statistical and neural machine learning methods are applied, including the use of large language models that are open and do not pose copyright or privacy issues.

To explore the complexity and potential for reducing cost and workload in annotating patient data, we tested large Norwegian language models as annotators.

What is sentiment analysis?

The task of sentiment analysis is to (i) identify subjective opinions and attitudes expressed in text, (ii) detect whether the opinion has a positive or negative polarity, and (iii) identify the targets and holders of the opinion. Additionally, aspect-based analysis will (iv) map the identified targets to more general topic categories.

References

Mæhlum, P., Samuel, D., Norman, R. M., Jelin, E., Bjertnæs, Ø. A., Øvrelid, L., & Velldal, E. (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* (pp. 8–19). Torino, Italia: ELRA and ICCL.

Results

Following meetings between both the annotators and NLP practitioners, we observed an increase in agreement. For the overall datasets, the annotation round before discussions contained many sources of disagreement, shown by especially low IAA-scores between certain annotators. However, following discussions, we note (Table 1) that IAA-scores has markedly increased for almost all annotator pairs. The agreement was especially high for binary (positive vs. negative) polarity, with all annotators reaching high agreement ($k > 0.9$) (results not shown).

Table 1: Inter-annotator agreement scores on intensity polarity after discussions. Change in scores in parenthesis.

	A1	A2	A3	A4	A5	A6
A1	1.0					
A2	0.78 (-0.01)	1.0				
A3	0.77 (-0.01)	0.86 (+0.07)	1.0			
A4	0.78 (+0.19)	0.89 (+0.31)	0.87 (+0.26)	1.0		
A5	0.77 (-0.06)	0.84 (+0.04)	0.84 (+0.12)	0.85 (+0.25)	1.0	
A6	0.78 (+0.09)	0.84 (+0.14)	0.81 (+0.14)	0.85 (+0.35)	0.79 (+0.10)	1.0

We tested large Norwegian language models as annotators and found that neutral sentences were challenging to classify. The models performed worse than humans, with lower agreement between model outputs and human annotations compared to human-to-human agreement.

We observed that the patient feedback data differed in formality, syntax, and lexical diversity when compared to the available Norwegian review data.

Table 2: Artificial examples of polarity annotation.

ID	Comment	Pos	Neg
0-1	Am very pleased with my current GP.	+++	
0-2	Don't want to change.	++	
0-3	Little info about what is happening.		-
0-Summary	Overall comment	+++	-
1-1	Very restrictive concerning pain medication.		---
1-2	Employees are not welcoming, one person in particular.		--
1-3	It was important for me to be hospitalized.	++	
1-Summary	Overall comment	++	---

+/- = slight positive/negative
++/-- = standard positive/negative
+++/- = strong positive/negative

Conclusions

While patient feedback data differ from existing datasets for Norwegian sentiment analysis, IAA-scores are promising for future modeling tasks, surpassing earlier reported scores for review data. This indicates a potential for good results for patient feedback sentiment analysis in Norwegian.

Automated analysis of comments offers broad applications, contributing to efficiency and cost savings. Integrating patient feedback into provider-level feedback strengthens patient voice in healthcare evaluation and quality improvement.

More information about the study and contact information to the research team can be found through scanning the QR-code

