

FINE-TUNING ROBERTA FOR SENTIMENT ANALYSIS OF SUPPORT CONVERSATIONS WITH PRIVACY-PRESERVING ANONYMISATION

Vojtech Mahdal¹, Oldrich Trenz¹

¹Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/ecos2026_020

ABSTRACT

This article describes a prototype system for processing support-conversation text. The system accepts a text input, classifies sentiment into three classes (positive, neutral, negative), masks selected sensitive items such as names, e-mails, phone numbers, URLs, usernames and long numeric strings, and stores an anonymised record. The sentiment classifier is based on RoBERTa and was fine-tuned in three stages on public chat and dialogue datasets. Evaluation showed that the second stage reached the best score on its internal test split, but its accuracy on a separate combined test set fell from 83.41% to 70.06%. After a final fine-tuning stage on a smaller dataset with a higher share of neutral examples, the model reached 92.82% accuracy on 1,573 independent samples. The anonymisation layer is rule-based and is most suitable for structured patterns; names, especially sentence-initial capitalised words, remain the main limitation. The result is a working demonstration consisting of a FastAPI service, a JavaScript dashboard and an SQLite log, containerised with Docker and deployed on Hugging Face Spaces.

Keywords: sentiment analysis, RoBERTa, fine-tuning, anonymisation, GDPR, customer support

JEL Code: C88

1 INTRODUCTION

Customer-support conversations can form a substantial body of free text that organisations may want to analyse, for example to understand recurring problems, customer attitudes and the tone of communication. At the same time, such texts may contain personal data, including names, e-mail addresses or phone numbers. This paper therefore treats sentiment analysis and anonymisation as linked tasks: the system extracts a three-class sentiment label from a text input and masks selected sensitive items before the result is stored (Mahdal, 2026).



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license. <https://creativecommons.org/licenses/by-sa/4.0/deed.en>

The thesis behind this paper (Mahdal, 2026) sets a narrow, practical goal: take a piece of conversation text, judge its sentiment, strip the sensitive parts, and keep only what is safe to store. Sentiment was chosen as the first structured signal to extract because it is useful on its own and because it lets the result be checked against published three-class models. The work is meant as a teaching example of the whole chain, from picking a base model to deploying a running service, rather than as a finished commercial product.

1.1 Extracting structure from conversations

Pulling structured facts out of free text sits in the wider area of information extraction, which recent surveys frame increasingly around large language models (Xu *et al.*, 2024). Sentiment analysis is one of its older and better-understood branches: Liu (2012) treats it as the task of deciding the polarity that a span of text expresses. Conversational settings add their own twist, because turn-taking and short replies carry context that a single sentence loses (Mariani *et al.*, 2023). We kept three sentiment classes rather than a positive–negative split, so that messages with no clear stance — a large share of support traffic — have a label of their own.

The model family used here goes back to the transformer (Vaswani *et al.*, 2017) and to BERT-style pretraining. Among the available encoders, RoBERTa (a Robustly optimised BERT approach) drops the Next Sentence Prediction objective, switches to dynamic masking and trains on a much larger corpus, roughly 160 GB against 16 GB for BERT (Kotamraju, 2022). For a three-class polarity task that trade favours accuracy over the lighter, faster DistilBERT, which is why RoBERTa was selected. The Hugging Face Transformers library (Hugging Face, 2022) supplied both the pretrained weights and the training loop.

1.2 Anonymisation as a constraint, not an afterthought

Because the stored text must be GDPR-compliant, anonymisation is part of the pipeline rather than a later clean-up step. The field offers a spread of options, from rule-based masking to learned models, with recent benchmarks comparing the two on the same data (Asimopoulos *et al.*, 2024; Shen *et al.*, 2024). The thesis takes the simpler route: a set of regular expressions that replace e-mails, URLs, phone numbers, usernames, long digit strings and capitalised name sequences with placeholder tokens. This is deliberately modest, and Section 4 is honest about where it slips.

2 METHODOLOGY

The pipeline runs in Python and has four parts: load data, mask sensitive items, fine-tune the classifier, then serve and log results. Before any model sees the text, a `mask_sensitive` function applies the regular expressions, turning e-mails into `<EMAIL>`, URLs into `<URL>`, phone numbers into `<PHONE>`, usernames into `<USER>`, long digit sequences into `<NUMBER>` and runs of capitalised words into `<NAME>`. Labels were mapped to integers (negative 0, neutral 1, positive 2) and text was tokenised with the pretrained RoBERTa tokenizer using padding and truncation.

Data were split in a stratified way, with roughly 70–80 % for training, 10–20 % for validation and the rest for testing. Class weighting was added to push back against imbalance, which mattered a great deal here: the first dataset was 75.78 % neutral, so without weighting the model could score well by guessing neutral. Training used the Hugging Face Trainer with a learning rate of $2e-5$,

three epochs and weight decay of 0.1. The work ran on MetaCentrum (CESNET) JupyterHub with a 6-core CPU, 32 GB RAM and Nvidia A40 GPU.

The classifier was not trained once but in three stages, each on a different dataset, building on the previous weights:

Stage 1 used `chat-sentiment-analysis` (Hugging Face, about 28.8k rows), first plain and then with class weights. **Stage 2** added EmpatheticDialogues (about 25k dialogues, 32 emotions remapped to three classes). The **final stage** added the small Chat Sentiment Dataset (about 500 rows), brought in specifically because the model still read the neutral class poorly. Why the middle stage hurt rather than helped is the more interesting part of the story, and Section 3 reports it with the numbers.

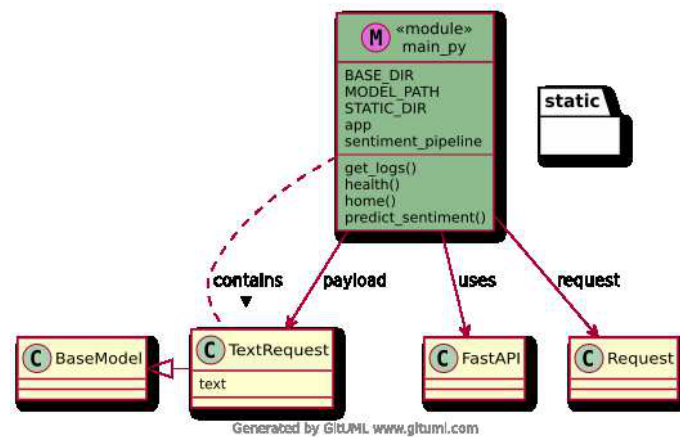


Fig. 1: UML class diagram of the main application module, showing the FastAPI service, the sentiment-analysis pipeline and the `predict_sentiment` endpoint (adopted from Mahdal, 2026)

The trained model was published to the Hugging Face Hub and loaded at runtime by a REST API built with FastAPI, which exposes a POST `/predict` endpoint and OpenAPI docs. A JavaScript dashboard sits on top of the API. Each prediction is written to an SQLite log through a thread-safe, regex-based anonymisation module, so the database never holds the raw sensitive text. The whole thing was containerised with Docker and deployed to Hugging Face Spaces. Figure 1 shows the class layout of the main module.

3 RESULTS

Two evaluation contexts have to be kept apart, because mixing them would misread the work. The first is the internal test split measured after each fine-tuning stage (Table 1). The second is a separate, combined test set of 1 573 unique samples that the model never saw during training (Table 2). The internal split and the independent set do not agree, and that disagreement is the point.

On the internal split the second stage looks like the winner, with an F1 of 0.9560. Taken at face value that would suggest stopping there. The independent set tells the opposite story.

Version 2, the one that scored highest on the internal split, dropped to 70.06% on the independent set, a fall of 13.35% against version 1 (Mahdal, 2026). The final version recovered and went past both, reaching 92.82%, which is an improvement of 22.76% over version 2 and the model shipped in the deployed service. The counts are consistent with the totals (1102 + 471 = 1573; 1460 + 113 = 1573).

The error breakdown in Table 3 shows where the trouble lived. Version 2 confused neutral with the two poles in 353 cases, far more than the other versions; almost all of its damage was neutral text pushed into positive or negative. The final version cut those mistakes sharply, and what remains is dominated by a single residual type, neutral read as positive (80 cases). Figure 2 shows the same picture as a confusion matrix for the final model.

Against two community RoBERTa sentiment models (`cardiffnlp/twitter-roberta-base-sentiment-latest` and `j-hartmann/sentiment-roberta-large-english-3-classes`) the trained model read standalone sentiments more accurately, the gap being clearest on the neutral class. This comparison is qualitative: the thesis presents it as side-by-side figures and does not report numeric accuracy for the two baselines, so none is quoted here.

The running service behaves as intended. On the dashboard the sentence “Everything is set up correctly!” was scored Positive with a confidence of 99.67 %, and a POST request with “FastAPI works correctly.” returned the matching positive label. Figure 3 shows the deployed interface.

Tab. 1 Metrics on the internal test split after each fine-tuning stage (not the independent set)

Metric	Stage 1 baseline	Stage 1 weighted	Stage 2 (dataset2v2)	Final (dataset3)
Eval loss	0.4481	0.5240	0.2030	0.3346
Accuracy	0.9243	0.9277	0.9561	0.9153
Precision	0.9282	0.9286	0.9560	0.9160
Recall	0.9243	0.9277	0.9561	0.9153
F1-score	0.9257	0.9281	0.9560	0.9152
Epochs	3	3	3	3

Tab. 2 Table 2: Accuracy on the combined independent test set (1 573 samples)

Model version	Accuracy	Correct	Incorrect
Version 1	83.41%	—	—
Version 2	70.06%	1102	471
Final version	92.82%	1460	113

Tab. 3 Most frequent error types by model version (count of class confusions)

Error (confusion)	Version 1	Version 2	Final version
neutral → positive	65	212	80
neutral → negative	54	141	19
positive → neutral	72	88	5
negative → neutral	68	26	7
positive → negative	2	3	1

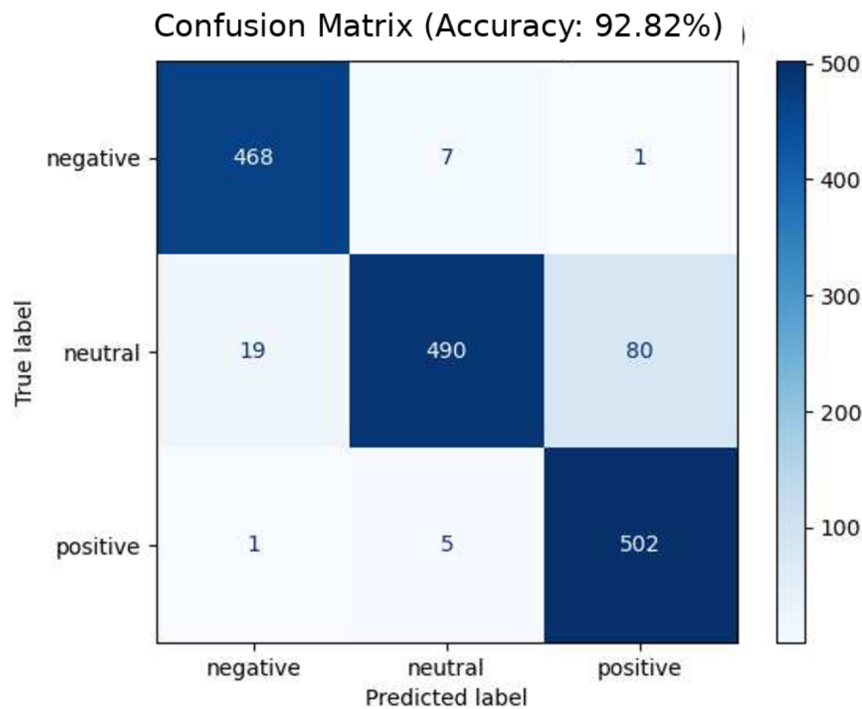


Fig. 2: Confusion matrix of the best-performing sentiment classification model on the three-class task (negative/neutral/positive), reaching 92.82% accuracy (adopted from Mahdal, 2026)

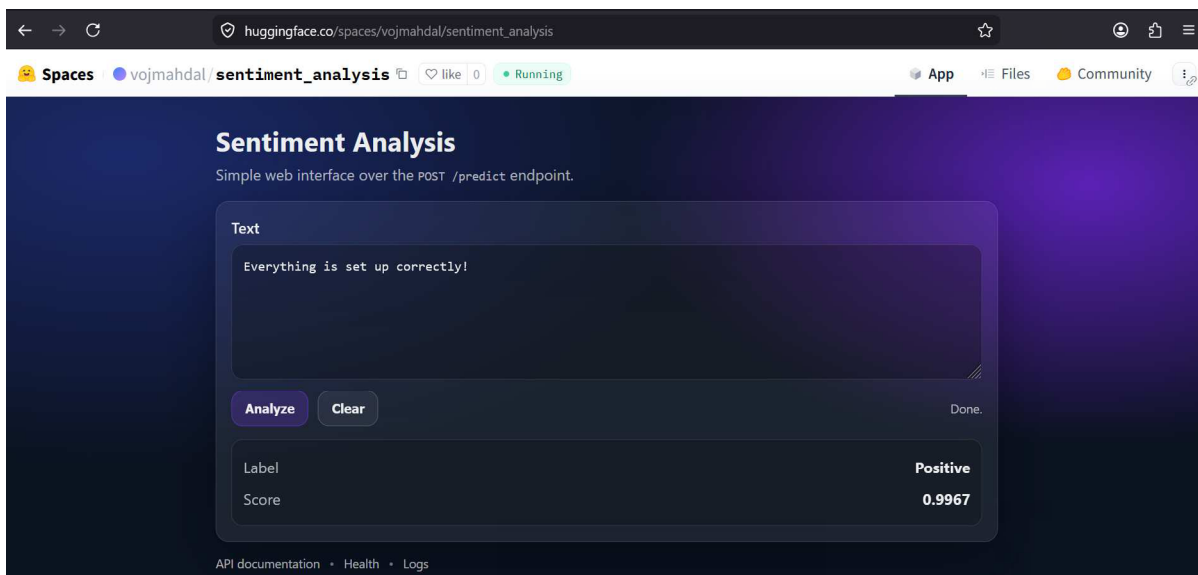


Fig. 3: Deployed web interface (Hugging Face Spaces) over the POST /predict endpoint, returning the predicted sentiment label and confidence score for an input sentence (adopted from Mahdal, 2026).

4 DISCUSSION

The gap between Table 1 and Table 2 is the main result worth discussing. The second stage achieved the strongest internal-split metrics, but the independent test set showed a much weaker result. The thesis links this behaviour to the different class distribution after remapping Empathetic Dialogues from 32 emotions into three sentiment classes, especially to weaker representation of neutral examples. It is therefore more accurate to describe this as a plausible explanation supported by the error pattern, not as a proven causal mechanism (Mahdal, 2026).

A plausible interpretation is that the final dataset helped mainly because of its class composition rather than its size. The Chat Sentiment Dataset contained a higher share of neutral examples than the preceding stage, and after fine-tuning on it the number of neutral-class confusions decreased from 353 in version 2 to 99 in the final version. The conclusion should therefore be formulated cautiously: in this experiment, adding a smaller but more balanced dataset improved generalisation on the independent test set, especially for the neutral class.

The anonymisation layer should be presented as a rule-based prototype rather than a fully reliable GDPR solution. Regular expressions are appropriate for structured patterns such as e-mails, URLs, phone numbers, usernames and long numeric strings, but the thesis itself identifies names as a difficult case, particularly when a capitalised word appears at the beginning of a sentence. For stronger anonymisation, the rule-based layer should be complemented or replaced by a trained named-entity recognition model, as suggested by recent anonymisation research (Asimopoulos *et al.*, 2024; Shen *et al.*, 2024).

One more caveat bounds the whole result. The independent test set was built from generated and public text made of short, plain sentences, so 92.82 % describes the model on simple, non-domain text. Sarcasm, irony, long compound sentences and the vocabulary of a specific product line were not tested. We would not expect the same figure on raw, messy support logs, and the work does not claim it.

5 CONCLUSION

The work produced a functional prototype for sentiment analysis and anonymised storage of support-conversation text. A RoBERTa classifier fine-tuned in three stages reached 92.82 % accuracy on a combined independent test set of 1,573 samples, and the model was deployed through a FastAPI service with a JavaScript dashboard and an SQLite log. The main transferable finding is that internal validation metrics alone were not sufficient: the second model version looked strongest on its internal split but performed poorly on the independent test set. The results suggest that the neutral class and class balance require special attention in three-class sentiment analysis. The main limitations are that anonymisation remains rule-based and incomplete for names, and that the reported accuracy was measured on generated and public short-text samples rather than on real support logs. Future work should therefore evaluate the system on domain-specific conversations and replace or supplement regex masking with a trained entity-recognition component.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-financed from the Operational Programme Research, Development and Education.

Anthropic Claude was used to summarize the text of the final thesis (Mahdal, 2026) and transform it into the form of this compact paper. The paper was then reviewed and fine-tuned by the author. The author is responsible for the correctness of the final text.

REFERENCES

- ASIMOPOULOS, D., SINIOSOGLU, I., ARGYRIOU, V., KARAMITSOU, T., FOUNTOUKIDIS, E., GOUDOS, S. K., MOSCHOLIOS, I. D., PSANNIS, K. E., SARIGIANNIDIS, P. 2024. Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches. Version 1. *arXiv:2404.14465* <https://doi.org/10.48550/ARXIV.2404.14465>
- HUGGING FACE. 2022. Transformers Documentation. *Hugging Face*. <https://huggingface.co/docs/transformers/v4.17.0/en/index>
- KOTAMRAJU, S. 2022. Everything You Need to Know About ALBERT, RoBERTa and DistilBERT. *Towards Data Science*. <https://towardsdatascience.com/everything-you-need-to-know-about-albert-roberta-and-distilbert-11a74334b2da/>
- LIU, B. 2012. *Sentiment Analysis and Opinion Mining*. San Rafael: Morgan & Claypool. <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
- MAHDAL, V. 2026. *Analysis and extraction of data from conversations between users and support services* [in Czech]. Master's thesis, head: Trenz, O. Brno: Mendel University in Brno. .
- MARIANI, M. M., HASHEMI, N., WIRTZ, J. 2023. Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda. *Journal of Business Research*. 161, 113838. <https://doi.org/10.1016/j.jbusres.2023.113838>
- SHEN, W. *et al.* 2024. *GDPR Compliance via NLP*. New York: ACM. <https://doi.org/10.1145/3690931.3691009>
- VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., POLOSUKHIN, I. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*. 30.
- XU, D. *et al.* 2024. Large Language Models for Generative Information Extraction: A Survey. *Frontiers of Computer Science*. 18, 186357. <https://doi.org/10.1007/s11704-024-40555-y>

Contact information

Vojtech Mahdal: e-mail: xmahdal@node.mendelu.cz

Oldrich Trenz: oldrich.trenz@mendelu.cz