Optimizing Context Use for Automated Classification of Teachers' Motivational Messages

EXPOSÉ FOR THE BACHELOR'S THESIS

Alexandre Koroneos
Humboldt-Universität zu Berlin

Supervision: Prof. Dr. Alan Akbik & Piet Wagner, M. Sc.

1 Introduction

Analyzing teachers' motivational messages in classroom discourse provides valuable insights into pedagogical practices and their impact on student engagement. However, manual annotation of such messages is resource-intensive and time-consuming (Bueno et al., 2025), limiting the scale at which classroom interactions can be studied. To address this challenge, the field has recently started to investigate the potential of automated classification using fine-tuned language models (Metzner et al., 2025a).

A critical factor influencing classification performance is conversational context. As illustrated by Figure 1, an utterance that appears ambiguous in isolation sometimes becomes interpretable when the preceding uterrances are known. Prior work has shown that incorporating this context improves model accuracy for dialogue act classification and related tasks in educational settings (Bueno et al., 2025). Despite this, the use of context in existing approaches remains largely unsystematic. Studies either provide no explicit justification for their chosen context window size or omit context entirely, leaving researchers without clear guidance on how to optimally leverage contextual information for their specific classification tasks.

This thesis addresses this gap by investigating the role of context in automated classification of teachers' motivational messages. Specifically, we examine three key questions: (1) What is the optimal amount of previous conversational context before performance plateaus or degrades? (2) Does including subsequent utterances improve classification accuracy? (3) To what extent does the optimal context window size differ between different transformer-based architectures?

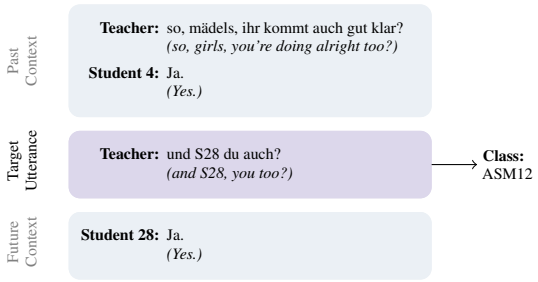

Figure 1: Example of a motivational message from the TALIS dataset. The target utterance "and S28, you too?" is ambiguous in isolation, but the context clarifies that the teacher is inquiring about the student's learning experience (class ASM12).

2 Related Work

2.1 Automated Analysis of Classroom Discourse

Manual annotation of classroom discourse is notoriously resource-intensive which limits the feasibility of large-scale studies. Recent work has therefore turned to natural language processing (NLP) to automate this classification (Metzner et al., 2025a). However, classroom discourse presents distinct challenges for NLP. Transcripts lack non-verbal cues like gestures and tone, and utterances often depend heavily on conversational context for their meaning.

Some researchers have explored multi-modal language models to capture the non-verbal dimensions, however, the question of how to present this information to language models in combination with verbal information remains a challenge (Bueno et al., 2025). The role of conversational context in the literature is explored in Section 2.3.

This thesis builds on Metzner et al. (2025b), who showed that transformer-based models can effectively classify motivational messages in teacher speech.

2.2 Transformer Architectures and Fine-Tuning Techniques

While large-scale decoder models (LLMs) have gained prominence for text generation tasks, for text classification, encoder-only architectures remain the more popular choice, due to their efficiency and performance (Benayas et al., 2025). In classroom discourse analysis, encoder models like DeBERTaV3 (He et al., 2023) serve as strong baselines. Bueno et al. (2025) found that domain-specific fine-tuning of these models outperforms prompting-based approaches, and Wang and Chen (2025) observed similar results with BERT (Devlin et al., 2019).

Despite this, some recent work has applied decoder models to classification tasks. Metzner et al. (2025a) reviewed several studies using LLMs for classifying teachers' motivational messages, some of which using in-context learning (ICL) and some fine-tuning approaches. Metzner et al. (2025b) successfully fine-tuned the decoder-only Gemma 2 (Gemma Team, 2024) for this purpose.

A parallel development is the optimization of decoder-only models for representation learning, which leverages the massive pre-training LLMs typically receive to generate dense embeddings for classification tasks (Tao et al., 2025). An example for this is the Qwen3-Embedding series (Zhang et al., 2025); these embedding-optimized decoders rank highly on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023), including on classification benchmarks.

2.3 The Role of Context in Dialogue Classification

As stated before, conversational context is critical for analyzing classroom discourse. In a study done by Bueno et al. (2025) where DeBERTaV3 was fine-tuned to label classroom activity sentence-by-sentence, performance improved significantly when the two preceding sentences were provided in addition to the target sentence. While the authors claim that "excessively long" context windows may be detrimental, it is not apparent whether they tested this and at which point a potential degradation may occur. Even if the context window was chosen experimentally, it is unclear to what extent this would have been determined by the data and task at hand or the model's capacity to handle the total amount of input.

Wang and Chen (2025) prepended one sentence to each target utterance when classifying dialogic moves with BERT, but did not test other context window sizes.

Two recent papers on classifying teachers' motivational messages, Falcon and Leon (2024) and Metzner et al. (2025b), used no conversational context at all. Metzner et al. (2025b) acknowledge that this might explain discrepancies between human annotations (which had access to context) and model predictions.

None of these studies provided future context to the model. However, a speaker's intent can sometimes become clearer from their next utterance or from the response they receive. Qamar et al. (2025) included future context in their work on dialogue act classification, though they worked with non-classroom data and did not directly compare setups with and without future context.

2.4 Handling Class Imbalance

Preliminary analysis suggests that class imbalance will be a significant issue in the TALIS dataset.

Metzner et al. (2025b) addressed this by using a class-weighted loss function based on the Effective Number of Samples (Cui et al., 2019), which scales the loss inversely to class density and prevents the model from simply predicting the majority class. They also attempted data augmentation with mixed results.

3 Research Questions

This thesis is conducted in the context of a research project at the University of Potsdam that seeks to validate and refine automated classification methods for teacher motivational messages, as introduced by Metzner et al. (2025b). While the broader project addresses multiple aspects of these classification methods, this thesis specifically focuses on investigating the role of conversational context as a potential driver of classification performance.

To this end, we address the following research questions:

RQ1: What is the ideal amount of previous conversational context for classifying teachers' motivational messages and at what point does additional context lead to performance degradation?

RQ2: Does the inclusion of subsequent utterances (future context) improve classification accuracy?

RQ3: To what extent does optimal context size

differ across the three dominant approaches for dialogue classification: smaller encoders, generative LLMs, and embedding-optimized LLMs?

4 Data

We use classroom discourse data from the TALIS Video Study Germany (Klieme et al., 2019), available through the FDZ Bildung (Research Data Centre Education). The dataset contains 138 video-recorded mathematics lessons from secondary schools (*Sekundarstufe 1*), all on the topic of quadratic equations. The videos have been manually transcribed, providing the basis for our work on automated classification of teachers' motivational messages.

The transcripts went through several preprocessing steps. After the initial manual transcription by the TALIS study team, our research group used the Deep Punctuation model (Guhr et al., 2021) for sentence segmentation and separated utterances at sentence boundaries. Student assistants then manually validated the transcripts to fix speaker attribution errors and other inaccuracies. Later, during annotation, the annotators adjusted punctuation and sentence boundaries where needed to improve segmentation quality.

We are working with manually annotated data from this corpus. Three trained annotators are currently annotating the transcripts using two classification systems: Autonomy-Supportive Messages (14 categories) and Value Messages (10 categories). The process started with all three annotators working together to establish inter-annotator agreement. After this phase, most transcripts will be annotated individually. We expect to receive about 20 % of the annotated transcripts (28 out of 138) by the end of January 2026, with more coming in mid-February for a total of roughly 45 transcripts. Additional annotated material should arrive by early March when we begin experiments.

We may adjust the final classification taxonomy based on how frequently categories appear and on inter-annotator agreement. Following Metzner et al. (2025b), we might remove very rare categories or merge related ones to make the classification system more robust.

5 Methods and Evaluation

To answer our research questions, we will fine-tune several pre-trained language models on the TALIS dataset and compare to which degree each model benefits from conversational context.

5.1 Model Architecture

To test whether the optimal context window size varies between model architectures (RQ3), we will compare three categories of models:

1. **Encoder-only models:** Following Bueno et al. (2025), DeBERTaV3 (He et al., 2023) serves as our primary encoder baseline, as it demonstrated strong performance in their classroom activity classification experiments. If preliminary tests indicate superior performance, we may use its multilingual variant mDeBERTaV3 instead. To test whether our findings generalize beyond the DeBERTa architecture, we include ModernGBERT (Wunderle et al., 2025), a German-specific encoder model.

2. **Large-scale decoder models (LLMs):** As our decoder baseline, we use Gemma 2 (Gemma Team, 2024), which achieved strong performance in Metzner et al.'s (2025b) classification of teachers' motivational messages. We also include Qwen3 (Yang et al., 2025) as a more recent decoder model.

3. **Embedding-optimized LLMs:** Qwen3-Embedding (Zhang et al., 2025) will serve as a representative for the approach of optimizing large-scale decoders for representation learning. It currently achieves the best performance on German classification tasks on MTEB (Muennighoff et al., 2023).

5.2 Fine-tuning

All models will be initialized with pre-trained weights from Hugging Face (Wolf et al., 2020) and fine-tuned on the TALIS dataset. For the larger decoder-based models, we will use Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) (Hu et al., 2021), similar to Metzner et al. (2025b). Hyperparameters will be adopted from prior work rather than performing dedicated hyperparameter optimization.

We will incorporate context into the models' inputs by prepending the k preceding utterances $u_{n-k}, ..., u_{n-1}$ to the target message u_n and separate them with special tokens, which we add to the tokenizer (cf. Bueno et al., 2025).

To handle class imbalance, we will use an imbalance-aware loss function, such as Focal Loss (Lin et al., 2018) or Weighted Cross-Entropy Loss

RQ	Variable	Configurations
1	Past Context Window	$k \in \{0, 1, 2, 4\}$ (Past utterances only)
2	Future Context	Shifted Window: $k_{\text{past}} = k_{\text{opt}} - 1$, $k_{\text{fut}} = 1$ (Total length $= k_{\text{opt}}$) Extended Window: $k_{\text{past}} = k_{\text{opt}}$, $k_{\text{fut}} = 1$ (Total length $= k_{\text{opt}} + 1$)
3	Model Architecture	Encoders: DeBERTaV3, ModernGBERT Decoders: Gemma 2, Qwen3, Qwen3-Embedding

Table 1: Overview of the planned experiments

with weights proportional to the inverse Effective Number of Samples (Cui et al., 2019). The exact choice will be determined through preliminary tests.

5.3 Experimental Design

Table 1 shows an overview of our planned experiments. For each model, we will first establish a baseline using only the target utterance ($k = 0$), then systematically test the impact of conversational context:

Experiment 1 (RQ1): To find the optimal past context window, we will test configurations with $k \in \{1, 2, 4\}$ preceding utterances. The goal is to find the point where the model captures relevant discourse patterns before hitting diminishing returns or degradation from noise. If a model does not hit this point at $k = 4$, we may expand the search space accordingly.

Experiment 2 (RQ2): To test whether future context helps, we will give the model one subsequent utterance in addition to past context. To keep the scope manageable, we restrict this to two configurations using the optimal window size k_{opt} from Experiment 1: a *shifted window* (keeping total length at k_{opt} by replacing one past utterance with one future utterance) and an *extended window* (adding one future utterance to all k_{opt} past utterances).

We address RQ3 by running the experiments above for each architecture described in Section 5.1.

5.4 Evaluation Strategy

We will use stratified k-fold cross-validation to reduce variance from rare classes. We will also implement grouped splits to prevent data contamination: all utterances from a single transcript stay strictly within either the training or test split. This prevents models from simply memorizing individual teachers' linguistic styles rather than learning the actual motivational categories. Without grouping, models might achieve high performance by recognizing a specific teacher rather than the desired patterns. We lack metadata to group by teacher across lessons, but grouping by transcript should work as a reasonable proxy to reduce speaker-specific bias.

Our primary metric will be the macro-averaged F1-score. We will also analyze class-specific performance to see whether certain experimental settings help some categories more than others. Beyond quantitative metrics, we may perform a brief qualitative error analysis, examining cases where the model failed without context but succeeded with it.

6 Scope and Limitations

This thesis focuses exclusively on fine-tuned transformer-based models for classification. Several related directions are outside our scope:

1. We do not investigate in-context learning (ICL) with LLMs. While ICL works well in low-resource settings (Brown et al., 2020) and could exploit the large context windows of modern LLMs, fine-tuning approaches appear to outperform ICL for our task and domain (see Section 2.2).

2. We do not use multi-modal approaches that combine audio or video data with transcripts. Multi-modal analysis has shown promising results for classroom discourse (Bueno et al., 2025), but such extensions are beyond the scope of this thesis.

3. We do not examine long-range dependencies beyond the immediate conversational context. It is worth noting that human annotators likely rely on such broader context when making annotation decisions, and incorporating this information might be necessary to achieve human-level classification accuracy.

4. We initially planned to test whether models fine-tuned on one context window size

could generalize to different sizes at inference time. This could reduce training costs, but we dropped this direction due to time constraints.

7 Timetable

The proposed progression of this research is detailed in Table 2.

References

Alberto Benayas, Miguel Angel Sicilia, and Marçal Mora-Cantallops. 2025. A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance. *Language Resources and Evaluation*, 59(3):2007–2030.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Ivo Bueno, Ruikun Hou, Babette Bühler, Tim Fütterer, James Drimalla, Jonathan Kyle Foster, Peter Youngs, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2025. Exploring automated recognition of instructional activity and discourse from multimodal classroom data. *Preprint*, arXiv:2512.00087.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. *Preprint*, arXiv:1901.05555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Samuel Falcon and Jaime Leon. 2024. Towards an optimised evaluation of teachers' discourse: The case of engaging messages. *Preprint*, arXiv:2412.14011.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2021. FullStop: Multilingual deep models for punctuation prediction. In *Proceedings of the Swiss Text Analytics Conference 2021*, Winterthur, Switzerland. CEUR Workshop Proceedings.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Eckhard Klieme, Juliane Grünkorn, Anna-Katharina Praetorius, Patrick Schreyer, Benjamin Herbert, and Julia Käfer. 2019. TALIS-Videostudie Deutschland.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Olivia Metzner, Yindong Wang, Gerard De Melo, Wendy Symes, Yizhen Huang, and Rebecca Lazarides. 2025a. The potential and limitations of large language models for automatic classification of teachers' motivational messages in educational research. *British Journal of Educational Psychology*, page bjep.70013.

Olivia Metzner, Yindong Wang, Wendy Symes, Yizhen Huang, Lena Keller, Gerard De Melo, and Rebecca Lazarides. 2025b. A process-oriented perspective on pre-service teachers' self-efficacy and their motivational messages: Using large language models to classify teachers' speech. *British Journal of Educational Psychology*, 95(S1).

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. 2025. LLMs are also effective embedding models: An in-depth overview. *Preprint*, arXiv:2412.12591.

Deliang Wang and Gaowei Chen. 2025. Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy. *British Journal of Educational Technology*, 56(6):2671–2704.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New encoders for german trained from scratch: Comparing ModernGBERT with converted LLM2Vec models. *Preprint*, arXiv:2505.13136.

Phase	Timeframe	Milestones & Activities
1. Preparation	Until Jan 31	• Finalization of the exposé • Drafting of chapters *Introduction* and *Related Work*
2. Implementation	Feb 01 – Feb 28	• Implementation of the data processing pipeline • Validation of setup using dummy data
3. Experiments	Mar 01 – Mar 21	• Arrival of TALIS data • Execution of the planned experiments • Writing of chapter *Methods*
4. Evaluation	Mar 22 – Apr 15	• Quantitative and qualitative analysis of the results • Writing of chapters *Results* and *Discussion*
5. Submission	Apr 16 – Apr 30	• Final re-run of best-performing models on extended data • Writing of chapters *Conclusion* and *Abstract* • Proofreading and submission

Table 2: Planned progression of this thesis

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.