Computational Humanities Research

www.cambridge.org/chr

From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data

Jannis Klähn[1,2], Janos Borst-Graetz[1] and Manuel Burghardt[1]

[1]Leipzig University, Computational Humanities, Leipzig, Germany and [2]Saxon Academy of Sciences and Humanities, Leipzig, Germany

Research Article

Cite this article: Klähn Jannis, Janos Borst-Graetz and Manuel Burghardt. 2025. "From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data." *Computational Humanities Research*, **1**:e4, https://doi.org/10.1017/chr.2025.10005

Received: 31 January 2025
Revised: 24 April 2025
Accepted: 2 June 2025

Keywords:
dictionaries; evaluation; large language models (LLMs); sentiment analysis; zero shot

Corresponding author:
Jannis Klähn;
Email: jannis.klaehn@uni-leipzig.de

Abstract

In this study, we perform a comprehensive evaluation of sentiment classification for German language data using three different approaches: (1) dictionary-based methods, (2) fine-tuned transformer models such as *BERT* and XLM-T and (3) various large language models (LLMs) with zero-shot capabilities, including natural language inference models, Siamese models and dialog-based models. The evaluation considers a variety of German language datasets, including contemporary social media texts, product reviews and humanities datasets. Our results confirm that dictionary-based methods, while computationally efficient and interpretable, fall short in classification accuracy. Fine-tuned models offer strong performance, but require significant training data and computational resources. LLMs with zero-shot capabilities, particularly dialog-based models, demonstrate competitive performance, often rivaling fine-tuned models, while eliminating the need for task-specific training. However, challenges remain regarding non-determinism, prompt sensitivity and the high resource requirements of large LLMs. The results suggest that for sentiment analysis in the computational humanities, where non-English and historical language data are common, LLM-based zero-shot classification is a viable alternative to fine-tuned models and dictionaries. Nevertheless, model selection remains highly context-dependent, requiring careful consideration of trade-offs between accuracy, resource efficiency and transparency.

Plain Language Summary

Sentiment analysis is a method used to determine whether a piece of text expresses a positive, negative, or neutral opinion. Traditionally, researchers have used sentiment dictionaries – lists of words with assigned sentiment values – to analyse text. More recently, machine learning models have been developed to improve sentiment classification by learning patterns from large datasets. However, most research in this area has focused on English, while many languages – including German – are under-represented. This study examines different sentiment analysis techniques to determine which methods work best for analysing German-language texts, including historical documents. We compare three main approaches:

1. Dictionaries: A simple and transparent way to check words against predefined sentiment lists.
2. Fine-tuned transformer models: These models, such as *BERT*, are trained on sentiment data to improve classification accuracy.
3. Large Language Models (LLMs) with zero-shot capabilities: Newer deep learning models, such as ChatGPT, can classify sentiment without being explicitly trained for the task by using their broad general knowledge.

To test these methods, we used a variety of datasets, including social media posts, product reviews and historical texts. Our results show that dictionaries are easy to use but often inaccurate, especially for complex texts. We also show that fine-tuned models perform well but require a lot of labelled training data and computing power. Zero-shot models, especially dialogue-based models like *ChatGPT*, are very effective, even without training on sentiment data. However, they are sensitive to small changes in input instructions and require significant computational resources. Our findings indicate that LLM-based zero-shot classification is a promising tool for sentiment analysis in the computational humanities. It allows researchers to analyse texts in different languages and historical contexts without requiring large amounts of labelled data. However, choosing the right model still requires careful consideration of accuracy, accessibility and transparency, especially when dealing with complex datasets.

Introduction

Sentiment analysis is a key area of research within natural language processing that focuses on understanding the emotional tone, attitudes and evaluations expressed in text.

A widely used approach within this field is sentiment classification (Pang, Lee, and Vaithyanathan 2002), which is often considered synonymous with sentiment analysis itself. One popular and rather simplistic approach is polarity-based sentiment classification, where sentences or documents are assigned to predefined categories such as *positive* and *negative*. Sentiment analysis initially gained popularity in the study of user-generated content on the social web, such as social media posts and review texts and for commercial purposes (Liang et al. n.d.). However, it is now also used in the computational humanities for a wide range of research applications, including computational linguistics (Taboada 2016), computational literary studies (Dennerlein, Schmidt, and Wolff 2023; Kim and Klinger 2019; McGillivray 2021) and digital history (Borst et al. 2023; Sprugnoli et al. 2016).

In this article, we provide a comprehensive evaluation of different sentiment analysis techniques for the case of German language corpora. This choice of language is motivated by a previous research project on media sentiment,[1] which focused on sentiment analysis in historical German newspapers. Given that most reference datasets used for sentiment analysis evaluations are based on contemporary English (Jim et al. 2024), we believe that our evaluation study of methods with a focus on their performance on historical German texts provides a valuable contribution to the existing evaluation landscape. In particular, it enhances the field of computational humanities, where non-English historical languages are often the subject of research. For our evaluation, we focus on off-the-shelf methods and models, which means they can be acquired and applied to new target data without any kind of adaptation or, in the case of neural networks, fine-tuning.

A popular branch of off-the-shelf techniques are dictionary-based methods. These are essentially pre-defined lists of words and their specific sentiment values or scores, making them easy to interpret and implement. Dictionaries are computationally efficient, versatile and compatible with most hardware. However, their static nature limits their adaptability to evolving language trends and they often struggle with domain-specific terminology and more complex linguistic phenomena such as sarcasm or negation. Although specialised dictionaries can be developed without advanced technical skills, the process is very labour intensive. As an alternative to dictionaries, there are several machine learning approaches to sentiment analysis. The traditional supervised learning paradigm, which involves training models on labelled data, has advanced considerably with the emergence of transformer architectures and large language models (LLMs).[2] Notably, models such as *BERT* (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) have contributed significantly to the rise of deep learning methods.

While dictionary-based approaches are highly dependent on language-specific features and often struggle with issues like orthographic errors or unknown words, the tokenisation procedures used by recent LLMs address these challenges more effectively. For instance, rare or out-of-vocabulary words are often broken into smaller subword tokens that the model can still interpret based on its training. Additionally, the contextualisation capabilities of LLMs enables them to automatically recognise negations and capture more nuanced semantic relationships. However, despite substantial improvements in processing speed and accuracy, generating training data for supervised learning remains a time-intensive task, especially in humanities projects that involve complex language data.

A promising development in this regard can be found in a branch of techniques known as zero-shot learning (Yin et al. 2019), a form of transfer learning that eliminates the need for task-specific annotated data. Zero-shot models use general knowledge gained from pre-training, allowing them to adapt to new domains with minimal to no customisation. Although these approaches offer flexibility and scalability, they remain computationally intensive and can pose challenges in terms of interpretability. The emergence of powerful LLMs such as *GPT-3* (Brown et al. 2020) and its chatbot interface, *ChatGPT*, has fundamentally changed the way we interact with these models. Dialogue-based LLMs allow tasks to be expressed through natural language prompts, greatly extending their applicability to different domains (Kocoń et al. 2023). Beyond the accessibility of corporate solutions such as ChatGPT, concerns have emerged about security, resource consumption and cost. In response, a growing number of locally deployable, smaller and open dialogue LLMs have been developed, providing an alternative for individual use in different configurations.

Given the variety of available off-the-shelf approaches available for sentiment classification, ranging from dictionary-based methods to dialogue-based LLMs, it remains unclear how effectively these techniques perform and compare when applied to German historical text corpora. This study addresses this gap by evaluating a number of sentiment dictionaries, fine-tuned language models and different zero-shot approaches to sentiment analysis. We test the different approaches on a wide range of German-language datasets containing contemporary social media and review data as well as language samples from the humanities domain. These *humanities texts* are examples of more specific text collections from different disciplines, including history and literary studies. Compared to the contemporary datasets, the humanities datasets are characterised by a much higher degree of heterogeneity in terms of text length, language used and context of creation. With this evaluation study, we extend our previous research on sentiment analysis for German language corpora (Borst et al. 2023) by adding a wide range of LLM-based zero-shot classification techniques.[3] We provide an in-depth evaluation of their performance and explore the broader implications of their use in computational humanities, considering factors such as accuracy, efficiency, interpretability and practical applicability. The results of this evaluation are meant to guide researchers in selecting appropriate sentiment analysis methods for their research projects and to better understand the trade-offs of different approaches.

Related work

For a long time, dictionary-based sentiment analysis has been a lightweight and therefore popular approach (Kolb et al. 2022; Lee et al. 2022; Mengelkamp et al. 2022; Müller et al. 2022;

[1] For more information on the research project ‘More than a Feeling – Media sentiment as a mirror of investor's expectations at the Berlin stock exchange, 1872-1930,’ see http://media-sentiment.uni-leipzig.de/.

[2] As model development progresses and the number of parameters increases, the definition of LLMs continues to evolve. For the purposes of this article, we define an LLM as a model with at least one billion parameters, developed after the first generation of transformer models.

[3] This previous work was published as part of the proceedings of the Computational Humanities Research conference in 2023 (Paris). It has been substantially extended in terms of the techniques evaluated and the corresponding discussion of results and implications for sentiment analysis in computational humanities settings.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Pöferlein 2021; Puschmann et al. 2022; Schmidt et al. 2021.[4] However, a major criticism of this method is its strong dependence on domain-specific classification performance (Borst et al. 2023; van Atteveldt et al. 2021), which requires extensive revalidation to achieve satisfactory results (Chan et al. 2021). In addition, sentiment dictionaries are inherently language dependent and cannot be directly translated without verification due to lexical ambiguity. Hybrid methods that integrate machine learning with semi-automatic word list generation or dictionary expansion have been proposed as promising alternatives. However, these approaches are often cumbersome due to the multiple validation steps required (Dobbrick et al. 2022; Palmer et al. 2022; Stoll et al., 2023). While dictionaries offer a low-barrier and resource-efficient solution that does not require training data (Schmidt et al. 2021), they consistently underperform compared to supervised learning methods. This is true for both off-the-shelf and custom dictionaries, including self-implemented and commercial options (Barberá et al. 2021; Boukes et al. 2020; Dobbrick et al. 2022; van Atteveldt et al. 2021; Widmann and Wich 2022).

In supervised learning, the fine-tuning of transformer-based language models, such as *BERT* (Devlin et al. 2019), has become the de facto standard for text classification tasks (Liu et al. 2019; Yang et al. 2019). Traditionally, the technical development of new methods for sentiment classification is often centred around the English language. This disparity is reflected in the research literature. In the context of German sentiment classification, supervised training approaches using language models achieve significant improvements in classification performance and generally provide a reference benchmark for other approaches (Barbieri et al. 2022; Guhr et al. 2020; Idrissi-Yaghir et al. 2023; Manias et al. 2023). The ability to fine-tune language models to a specific grammatical or morphological context proves particularly successful for sentiment classification in historical German language (Borst et al. 2023; Schmidt et al. 2021).

A key challenge in applying language models to domain-specific tasks is the need for annotated training data as well as for substantial computational resources (Schwartz et al. 2019). Domain adaptation through fine-tuning typically requires updating millions of parameters for each dataset, which can be computationally expensive. To address these issues, recent research has focused on reducing the reliance on large labelled datasets, leading to the rise of few shot (Bao et al. 2020; Bragg et al. 2021; Brown et al. 2020; Wang et al. 2020) and even zero-shot models (Schönfeld et al. 2019; Xian et al. 2017; Yin et al. 2019). These approaches enable text classification without the need for extensive task-specific fine-tuning or manual data annotation, significantly lowering the barrier to entry.

An important milestone in this area was the introduction of *GPT-3* (Brown et al. 2020). Research into instruction-following models (Ouyang et al. 2022; Wei et al. 2021) and the publication of dialogue-based systems, such as *ChatGPT* or *Llama* (Touvron et al. 2023), have further increased accessibility. Initial research indicates strong zero-shot classification performance of *ChatGPT* (Gilardi et al. 2023; Törnberg 2023), although challenges such as non-determinism remain (Reiss 2023). Recent studies have increasingly explored their application to sentiment classification, particularly using OpenAI models as a comparison (Campregher and Diecke 2024; Jim et al. 2024; Kheiri and Karimi 2024; Rauchegger et al. 2024; Wu et al. 2024; Zhang et al. 2024; Zhu et al. 2024). However, these benchmarks focus on English language datasets.

One way of dealing with non-English languages is to rely on machine translation (Feldkamp et al. 2024; Koto et al. 2024; Miah et al. 2024), to make use of available tools for English. While a manual checking of translation quality remains a viable option for smaller datasets (Campregher and Diecke 2024), this may become a factor when moving to historical or literary texts, given the unique and more complex language setting (Etxaniz et al. 2024; Huang et al. 2023; Liu et al. 2025). Recently, a growing number of LLMs have been optimising multiple languages simultaneously. However, as English often continues to dominate the training data of these models, research shows that switching to non-English languages reduces the level of performance within these models (Etxaniz et al. 2024; Zhang et al. 2023). Notably, there are some examples for cross-lingual zero-shot approaches, which seem to offer more consistent sentiment classification performance in non-English languages (Koto et al. 2024; Manias et al. 2023; Přibáň and Steinberger 2022; Sarkar et al. 2019; Wu et al. 2024).

Off-the-shelf methods and models for the German language are still scarce, and even more so in the field of computational humanities data. Among the few are Guhr et al. (2020) and Barbieri et al. (2022), offering ready-to-use fine-tuned models for German sentiment analysis. However, these models are trained on contemporary Twitter or review datasets, raising the question of how well their performance transfers to out-of-domain data. So far, sentiment analysis studies on historical German texts have been performed using dictionaries (Du and Mellmann 2019; Schmidt and Burghardt 2018a; Schmidt et al. 2021) or machine learning methods, such as SVMs (Zehe et al. 2017) or fine-tuning (Borst et al. 2023; Dennerlein, Schmidt, and Wolff 2023; Schmidt et al. 2021). Yet, preliminary results from our previous work show that even a rather small German BERT-based zero-shot model can potentially deliver performance comparable to the aforementioned fine-tuned models on contemporary datasets and even outperform them and dictionary approaches on humanities datasets (Borst et al. 2023).

Methods and experiments

Our previous research (Borst et al. 2023) demonstrated that an natural language inference (NLI)-based zero-shot classifier consistently outperformed dictionary-based approaches across all datasets for polarity-based sentiment classification, although it did not achieve state-of-the-art (SotA) performance. Two key observations emerged from this study: (1) the zero-shot model demonstrated consistent performance patterns across both contemporary and humanities datasets, whereas fine-tuned and dictionary methods experienced significant performance declines on these datasets; and (2) the performance of dictionary-based approaches was notably inconsistent. These findings position the NLI-based zero-shot model as a promising middle ground between the efficiency of dictionary-based approaches, and the high performance of fine-tuned models, which come at a significant computational cost.

In this section, we extend the evaluation of the methods on the benchmark datasets of Borst et al. (2023) by including a broader range of sentiment classification techniques Table 1. We then provide an overview of the main classes of methods identified in the literature and justify our choice of models for evaluation. Finally,

[4] Disclaimer: This article cites preprints, such as those at https://arxiv.org/, to reflect the latest developments in the rapidly evolving field of LLMs. Preprints are preliminary reports of research and have not been peer-reviewed.

The purpose of including preprints is to discuss the current state of research. Readers are advised to interpret these sources with caution.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Table 1. Aggregated list of all evaluated LLMs, including their respective names as they appear on Hugging Face or version name

Model	URL / version	Size
germanSentiment	oliverguhr/german-sentiment-bert	109 M
XLM-T	cardiffnlp/twitter-xlm-roberta-base-sentiment	278 M
BERT	svalabs/gbert-large-zeroshot-nli	336 M
mDeBERTaX	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	279 M
MPNET-XNLI	sentence-transformers/paraphrase-multilingual-mpnet-base-v2	278 M
RoBERTaX	T-Systems-onsite/cross-en-de-roberta-sentence-transformer	278 M
RoBERTa-Sentence	T-Systems-onsite/german-roberta-sentence-transformer-v2	278 M
gpt-3.5-turbo	gpt-3.5-turbo-0125	-
gpt-4o	gpt-4o-2024-08-06	-
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct	8 B
Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct	70 B
Ministral-8B	mistralai/Ministral-8B-Instruct-2410	8 B
gemma-2-9b	google/gemma-2-9b-it	9 B
SauerkrautLM-9b	VAGOsolutions/SauerkrautLM-gemma-2-9b-it	9 B
Teuken-7B	openGPT-X/Teuken-7B-instruct-research-v0.4	7 B

Note: It includes information on their individual parameter size values. The number of parameters of the GPT models are not publicly known.

we give a brief description, including critical remarks, of the benchmark datasets used in this evaluation study Table 2.

Dictionaries

We adopt the dictionary selection from Borst et al. (2023), which includes three widely used and universally applicable German sentiment dictionaries: *BAWL-R* (Võ et al. 2009), *SentiWS* (Remus et al. 2010) and *GermanPolarityClues* (GPC) (Waltinger 2010). To account for domain-specific variations, we also included the finance-specific dictionary *BPW* (Bannier et al. 2019) and the literary studies dictionary *SentiLitKrit* (SLK) (Du and Mellmann 2019) to also ensure coverage of humanities datasets. As the 'death of the dictionary' has already been claimed in Borst et al. (2023), it was decided not to include additional dictionary resources in the evaluation. Instead, we focus on extending our experiments by incorporating additional fine-tuned language models and, in particular, exploring a wide range of LLMs with zero-shot capabilities.

Fine-tuned language models

Task-specific, fine-tuned language models – often regarded as the precursors of modern LLMs – have proven to be a viable approach for sentiment analysis (Wang et al. 2023). The fine-tuning of a *BERT*-based language model to the target data still achieves the highest performance for various application areas, including almost all SotA results for the datasets we tested, and has long served as a benchmark.[5] Off-the-shelf models with task- or domain-specific training offer a wide range of applications even without the availability of annotated data.

We extend the model selection of our previous study (Borst et al. 2023), where the German multi-domain model *germanSentiment* (Guhr et al. 2020) was tested as a representative of the fine-tuned models, by adding the multilingual *RoBERTa*-based model, *XML-T* (Barbieri et al. 2022). This model, fine-tuned on millions of tweets for various tasks in different languages, provides an additional perspective on domain transfer of sentiment classification models. It is noteworthy that the training data of the fine-tuned models used in this evaluation overlaps with our benchmark Twitter and review datasets. However, the comparison still serves to answer two important questions: How well do the zero-shot methods perform on these datasets even though the fine-tuned models have seen them during training? How well does the performance of these models trained on contemporary data transfer to texts from the humanities domain? This is of particular interest, as these models may yet provide a standard method for performing sentiment analysis in the computational humanities.

When comparing our results with *germanSentiment*, we must include a disclaimer, as we were unable to replicate the exact test sets used in their reported results. In fact, competing versions of the models, each using different pre-processing methods, result in slight variations in performance values. Although we attempted to follow the authors' instructions for applying the models via Hugging Face, we were unable to reproduce the reported SotA results for any of the models. Therefore, for this particular model, our evaluation may differ from the values originally reported by the authors. Currently, the only sentiment classification models in German that have been fine-tuned for broader applicability to humanities texts are the two that were reported in this section, with no other ready-to-use models available. However, the models we tested provide insight into their performance transfer across different datasets and domains.

LLM-based zero-shot text classification

To systematise our evaluation study, we identified three popular categories of methods for LLM-based zero-shot text classification

[5] This is the case for all datasets except GND, Lessing and SLK.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Table 2. Statistics of all datasets used, including sentiment label distribution, average text length and temporal coverage

	Dataset	Negative	Neutral	Positive	Total	Avg. words	Std. dev.	Years
Humanities	BBZ Gold	260	198	314	772	23.24	16.33	1872–1930
	Lessing	139	-	61	200	53.32	42.13	1747–1779
	SentiLitKrit	292	-	718	1,010	37.61	22.43	1870–1899
	GND	89	124	57	270	16.99	7.77	
Twitter	GermEval	780	1,681	105	2,566	71.51	208.81	2015–2016
	PotTs	1,569	2,487	3,448	7,504	18.06	5.96	2013
	SB10k	1,130	4,629	1,717	7,476	14.57	8.43	2013
Review	Amazon	2,000	1,000	2,000	5,000	32.97	30.05	2015–2019
	Filmstarts	15,608	-	40,012	55,620	123.14	149.20	2018
	Holiday Check	11,099	-	88,901	100,000	61.30	69.02	2018
	SCARE	26,903	-	73,097	100,000	13.11	15.79	2014–2015

in the existing literature. Each category is represented in the experiments by a specific selection of pre-trained models.

Sentence pair classification

Sentence pair classification models are one possible approach to zero-shot text classification. The goal of sentence pair classification is to determine the relationship between two input sentences. Popular sentence pair tasks include *NLI*, also called *entailment*, or *next sentence prediction* (NSP). NSP assesses whether one sentence is likely to follow another in a text. In the case of NLI, the aim is to decide whether the second sentence logically entails or contradicts the first.

In our evaluation study, we use NLI as the reference method, as originally proposed in Yin et al. (2019). In this approach a sentence pairs, called premise and hypothesis, are classified as 'entailment,' 'contradiction' or 'neutral,' based on how well the hypothesis logically entails the premise. For zero-shot classification, we formulate hypotheses using the target labels. These hypotheses are created using the hypothesis template: '*The sentiment is* [*blank*]'.[6] The blank is then filled with the sentiment categories *negative*, *neutral* and *positive*. For application purposes the hypothesis template can have substantial impact on the quality of classification, and is part of the optimisation process similar to prompt engineering (Liu et al. 2023). Since we aim at comparing these models with zero knowledge about domain-specific assumptions or vocabulary, we use the same for all datasets and models. Each model generates probability scores for each premise and hypothesis pair, corresponding to the different entailment classes. From these scores, we identify the hypothesis with the highest probability of 'entailment' as the classification outcome and assign the corresponding category. Although there is some criticism about the performance of these models, particularly their reliance on spurious correlations in superficial text elements (Ma et al. 2021), these models (and their variants) still perform very well, especially in sentiment classification (Shu et al. 2022; Zhang et al. 2023). In Borst et al. (2023) a pre-trained *BERT*-based model was used. For our evaluation study, we extend the model selection of our previous study with a multilingual *mDeBERTa*-based model.

Similarity-based and Siamese networks

Methods in this category use embeddings to jointly embed text and labels into the same semantic space. By applying a similarity function (e.g., cosine similarity), the embeddings of labels and text are compared, and the resulting scores determine the label with the highest score. Originally proposed by Socher et al. (2013) and Veeranna et al. (2016), this approach has also been adopted in more recent studies (Mueller and Dredze, 2021; Molnar 2022). The key advantage of similarity-based methods is that they do not require explicit training on labelled sentiment datasets. Instead, they rely on pre-trained sentence encoders that capture general semantic relationships, making them applicable in a zero-shot setting. The model does not need to be fine-tuned on task-specific data. Rather, it assigns labels by measuring the similarity between input text and predefined class labels embedded in the same vector space.

Recently, pre-trained LLMs have been chosen as the backbone for the Siamese approach, where two identical neural networks process text and labels separately but share parameters to ensure a unified embedding space. Sentence-*BERT* (Reimers and Gurevych 2019) models present a viable option for this approach. These are a class of embedding models specifically fine-tuned to embed sentences and have significantly improved performance on the semantic textual similarity (STS) benchmarks (Cer et al. 2017). The final selection includes the cross-lingual MPNET-XNLI and two RoBERTa-based models fine-tuned on STS, namely RoBERTaX and RoBERTa-Sentence.

Instruction-following models or dialogue-based systems

In addition to using the generative capabilities of LLMs to predict individual tokens, instruction-following models (Wei et al. 2021; Ouyang et al. 2022) and dialogue-based systems (Peng et al. 2022; Zhang et al. 2020), such as *ChatGPT*, have made powerful models accessible to a wider audience. By instructing these models to generate a label from a set of pre-defined classes based on a task description and text input, they can effectively act as zero-shot text classifiers. Initial evaluations of zero-shot classification performance show promising results (Gilardi et al. 2023; Törnberg 2023), although they also highlight the caveat of non-determinism, such as the influence of the temperature hyperparameter (Reiss 2023). A critical factor influencing performance in this context, is the formulation of the task description (White et al. 2023). Additionally, models of this size typically cannot be run on conventional local hardware. This limitation has spurred a trend towards smaller

[6] Translated from German: '*Die Stimmung ist* [*Platzhalter*].'

instruction-following models, such as Llama (Touvron et al. 2023) and Mistral (Jiang et al. 2023), making dialogue-based models a viable alternative for locally executed classification tasks.

In selecting the models to test as a local alternative to industry leader GPT, particular attention was paid to technical compatibility. Specifically, the models had to run on an enterprise-grade NVIDIA A30 graphics card without quantisation. Quantisation is a technique for significantly reducing the hardware requirements of LLMs, but there is still a lack of sufficiently generalised understanding of its differential impact on different model families. Although recent studies suggest that moderate quantisation can be applied without significant performance impact (Jin et al. 2024; Liu et al. 2024), the choice of a quantisation approach and the desired level of precision introduce additional experimental variables. Furthermore, the availability of an executable model variant on the Hugging Face platform (Wolf et al. 2020) was considered essential to ensure potential applicability in humanities research. Although there are alternative solutions for running LLMs locally, such as *Ollama*[7] or *llama.cpp*,[8] Hugging Face offers a wide range of models that can be easily integrated and used in a straightforward manner. We place particular emphasis on models within the 7B to 9B parameter range, as this strikes a balance between memory consumption, inference time and performance. These models can also be used on conventional GPUs, allowing them to be run locally without the need for a dedicated high performance computing cluster.

The final selection includes *gpt-3.5-turbo-0125* and *gpt-4o-2024-08-06*, as well as *LLama-3.1* (Touvron et al. 2023) in its 8B and 70B configurations and *Ministral* 8B. These models are trained on multilingual data that also includes German. In addition, recent developments in specialised instruction models were included to examine whether language specificity significantly affects the results, including *SauerkrautLM* (a German Gemma variant) and its English-only base model *Gemma 2 9B* as well as*Teuken 7B*, a project aimed at an optimised model for all EU languages, co-funded by the German government.

All local LLMs were tested in the same test setup with the same prompt. The text was not pre-processed, as in German, for example, the removal of punctuation marks can significantly alter the semantic context. By using the Hugging Face pipeline, the results could be transferred directly as text output in separate columns in the datasets, which reduces possible formatting problems, but could be a possible factor in terms of speed or performance. The quality of the output was highly sensitive to the choice of the prompt and the chosen parameters, a factor that should not be underestimated and is briefly discussed here.

The question of which prompt produces the best results is subject to constant change, as research into the use of dialogue-based LLMs progresses and the models themselves are continually improved. While the initial research on sentiment analysis with LLMs was based on simple, reduced prompts, which were thought to have the greatest potential (Kheiri and Karimi 2024; Miah et al. 2024; Wu et al. 2024; Zhang et al. 2024), there are now numerous prompting strategies with sometimes very contradictory results. The possibility of enriching prompts with semantic context or exploiting the reasoning capacities of the models via a so-called *chain-of-thought* (COT), as well as the combination with few-shot approaches, did not necessarily lead to better results (Rauchegger et al. 2024; Wang and Luo 2023; Wu et al. 2024; Zhang et al. 2024). Furthermore, it is beyond the scope of this work to include multiple parameters to our broad model evaluation.

After some preliminary tests with a prompt as suggested by Kheiri and Karimi (2024) and a deterministic temperature setting, there were some significant deviations in the output. A more heavily formatted prompt based on current OpenAI recommendations was tested,[9] as well as various settings for the temperature. The combination of a temperature of `0.1` and the following prompt gave the highest consistency in the results over the entire test setup and was therefore adopted for all dialogue-based models:

```
TASK: Sentiment Classification
INSTRUCTION:Classify the following text into exactly
one sentiment category.
INPUT TEXT:
"{text}"

RULES:
- You must choose exactly ONE option: {labels}
- Respond with ONLY the chosen word
- DO NOT add any explanation or additional text
- DO NOT use punctuation or formatting

CLASSIFICATION:
```

Although this setup performed well in our tests, we refrain from making a generalised judgement due to the many additional influencing factors. In a non-deterministic setup, model behaviour may vary between different runs. This was investigated in preliminary tests, but no significant variations were observed. Despite the large context size of LLMs, the concatenation of multiple prompts per call introduced another source of error, leading us to adopt a slower, sequential inference approach. Nevertheless, incorrect outputs were observed in some cases, although they occurred least frequently in the chosen configuration. In this configuration, the only deviation from the expected labels was an empty output, which was considered an error and taken into account in the evaluation.

German language datasets

All the datasets used in this evaluation study were selected in the basis of open availability and mention of SotA results in recent research publications. Nevertheless, the availability of non-English datasets remains a major limiting factor and also poses a significant problem for the subsequent training of specialised language models. Following the evaluation design in (Borst et al. 2023), we used datasets from three different domains.

Humanities datasets

In addition to a total of seven contemporary German-language datasets that are based on social media posts and reviews, we also selected four domain-specific datasets from the humanties domain with a focus on historical German, which we henceforth refer to as *Humanities Datasets*. Based on previous research (Borst et al. 2023), the *BBZ* dataset (Wehrheim et al. 2023) was created from articles published between 1872 and 1930 in the Berliner Börsenzeitung, a stock exchange newspaper. The dataset was annotated by a domain expert and contains polarity-based sentiment annotations for a total of 772 sentences.

The *Lessing* dataset (Schmidt et al. 2018) is another example of historical text analysis, comprising 200 speeches from Gotthold

[7] https://ollama.com/

[8] https://github.com/ggerganov/llama.cpp

[9] https://platform.openai.com/docs/guides/prompt-engineering

Ephraim Lessing's theatre plays. These texts were annotated by five individual annotators and a domain expert, using binary sentiment labels. A similar approach was applied in the *SentiLit–Krit* (SLK) dataset (Du and Mellmann 2019), which consists of German literary criticism extracted from historical newspapers. This dataset includes 1,010 binary annotated sentences and originates from the same time period as the BBZ corpus. A broader time span is covered in the *German Novel Dataset* (GND) (Zehe et al. 2017), which was derived from the German Novel Corpus through crowdsourced annotation. It contains 270 sentences from various literary works, offering insights into sentiment analysis in a wider historical context.

Twitter datasets

The *GermEval 2017* dataset (Wojatzki et al. 2017) is based on tweets and other social media posts related to Deutsche Bahn and was compiled between 2015 and 2016 as a benchmark dataset for various sentiment-related tasks. The dataset was cleaned, manually annotated and sampled for evaluation. We use the predefined synchronous test set with 2,566 examples labelled with positive, neutral or negative values. The *PotTs* (Sidarenka 2016) dataset contains 7,504 items from tweets during the 2013 German federal election and other political topics. The *SB10k* (Cieliebak et al. 2017) dataset contains tweets from 2013 and was created with the intention of creating a German reference dataset. In a first step, the tweets were clustered to achieve a broad coverage of topics. In a second step, the contained polarity words were compared with the German Polarity Clues Lexicon (Waltinger 2010) to ensure that actual sentiments were included. These clusters were then balance-sampled and manually annotated. However, as the published dataset only includes the Twitter link and annotation, it was decided to use the version created by Guhr et al. (2020), which includes all tweets as text and the three labels positive, neutral and negative. *SB10k* is also often used as a German Twitter benchmark dataset in multilingual sentiment analysis experiments (Barbieri et al. 2022).

Review datasets

Datasets consisting of reviews are often used for sentiment analysis evaluation, because they are widely available and easily accessible. Many studies follow the approach of Pang, Lee, and Vaithyanathan (2002), who categorise reviews as *positive* or *negative* based on their star ratings (Guhr et al. 2020; Manias et al. 2023). For the sake of consistency, we used the same labelling strategy for our evaluation study as was used for the SotA results. This means that in most cases, 3-star reviews were excluded due to the difficulty in assigning a clear *positive* or *negative* label.[10]

The *Amazon Review* (Keung et al. 2020) dataset is based on product reviews from 2015 to 2019 and includes multiple languages with equal proportions of entries. The pre-defined test set of 5,000 items was selected for the German language. The *Filmstarts* (FS) and *Holidaycheck* (HC) datasets were both created by Guhr et al. (2020) for general sentiment classification in the German language and contain film and hotel reviews. The resulting datasets contain 55,620 entries for *Filmstarts* and almost 3,3 million entries for *Holidaycheck*. Furthermore, the *SCARE* (Sänger et al., 2016) dataset was selected, as it contains around 735,000 German-language reviews of almost 150 selected apps from the Google Playstore. The largest datasets, *Holidaycheck* and *SCARE*, were each sampled at 100,000 entries to reduce the significant inference time. An overview of the scope of the data records, a breakdown by represented labels and the average number of words per text in the data set, their standard deviation, and the covered time span in years is shown in Table 2.

Conclusive remarks on datasets

Regardless of domain and scope, the datasets vary in quality, which has a direct impact on the difficulty of classification. In addition, different annotation methods introduce a certain bias, as they shape the expectations about the nature of the data and its sentiment distribution. An important issue that often leads to problems with polarity-based sentiment detection is the ambiguity of the neutral label. This class can represent both mixed and non-sentiment statements and is not uniformly defined in its use. Especially in the case of reviews, the assumption that a mediocre rating is equivalent to neutral is problematic, so such ratings are often excluded (Guhr et al. 2020; Pang, Lee, and Vaithyanathan 2002). This decision is made primarily in the interest of optimising the classification metrics. In the context of humanities data, texts that are often ambiguous and fall between binary labels. Consequently, while the task is simplified by reducing the range of labels to binary options, the semantic quality of the classification itself and the applicability in more specific scenarios may be compromised. The resulting binary labels can usually be assigned more easily due to their diametric characters, which explains the extremely high classification results of all methods. Outside of this simple task, the SotA scores are significantly lower, indicating a higher level of difficulty.

However, a fundamental problem with user-generated content such as tweets is the large variation in text quality. While the *GermEval* dataset contains a large proportion of longer, official tweets with a clearer structure, the *PotTs* dataset contains a large number of semantically unclear texts in various samples. The assignment of these texts to a sentiment label seemed questionable. In contrast, the high SotA values on the *BBZ* data showed that even temporally and contingently extremely specific language contexts can be learned and used in the sentiment analysis task (Borst et al. 2023). Notably, the texts and annotations of the *GND* dataset proved to be significantly less accurate, as reflected in the poor performance of all the models tested so far on this dataset. This raises the question of whether a model that performs optimally on such mediocre data, or on a simplified task, will also perform well on other datasets or domains. As social media and review data are widely available, there is an imbalance in the optimisation of model development and optimisation in favour of these domains. This has to be taken into account for the specific questions in the humanities, which also lead to other data qualities and requirements.

Results

In presenting the results, we would like to emphasise that the aim of this evaluation study was primarily to assess how well different models can adapt to different application contexts, rather than to identify the optimal model for a particular dataset. The evaluation was carried out by comparing the results of the different approaches with the SotA values for the different datasets as they are reported in the literature. In addition, we were interested in analysing performance across different domains and model classes in order to obtain a balanced perspective on the trade-offs and capabilities involved in domain adaptation. Table 3 shows the micro-F1 evaluation scores for all datasets and approaches. Overall, the zero-shot models perform reasonably well, with some

[10] The Amazon reviews dataset is the only exception here, as the SotA results also considered 3-star ratings and – controversially – labelled them as *neutral*.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Table 3. This table presents the evaluation of all models and methods in this study based on micro F1 scores

	Humanities				Twitter			Reviews			
	BBZ	GND	Lessing	SLK	GermEval	PotTs	SB10k	Amazon	Filmstarts	Holidaycheck	SCARE
Dictionary											
SLK	37.1	49.6	38.7	66.2	48.5	38.8	53.9	42.5	67.4	64.9	31.5
BPW	43.5	47.4	55.7	51.5	56.3	40.6	48.7	46.4	59.6	69.6	36.8
BAWL-R	37.1	34.8	42.4	65.2	24.2	43.1	36.9	41.6	70.3	84.4	49.4
Senti-WS	51.9	43.7	58.2	66.5	38.0	46.1	36.5	58.1	74.3	85.3	72.2
GPC	51.1	45.5	60.8	62.1	36.6	43.7	43.5	58.2	71.9	82.4	72.6
Fine-tuned											
germanSentiment	27.2	44.8	50.6	11.1	58.3	38.9	61.4	66.9	83.1	93.5	79.7
XLM-T	38.8	48.1	62.9	37.0	58.3	**62.0**	**75.5**	66.6	73.8	78.2	79.5
NLI											
BERT	**67.6**	46.6	74.6	78.7	33.2	51.6	33.5	69.7	82.2	92.9	**87.9**
mDeBERTaX	67.4	45.6	76.0	75.5	33.1	47.7	30.7	35.6	82.5	80.9	85.4
Siamese											
MPNET-XNLI	54.5	44.1	61.0	78.8	50.1	46.0	54.1	24.1	73.2	61.2	84.9
RoBERTaX	56.7	46.3	61.5	57.8	49.2	44.1	47.9	23.8	68.9	73.8	79.4
RoBERTa-Sentence	51.8	46.3	66.0	52.2	54.0	44.0	53.0	22.2	66.0	66.1	79.1
Dialogue											
gpt-3.5-turbo	66.5	54.8	76.0	73.7	48.3	62.3	60.4	73.2	93.6	91.6	86.2
gpt-4o	57.3	**57.0**	79.5	71.3	62.4	56.6	70.9	75.0	95.8	92.7	86.9
Llama-3.1-8B	33.9	48.5	**81.3**	84.8	62.1	51.9	70.4	71.3	93.8	93.8	84.4
Llama-3.1-70B	65.5	53.5	**81.3**	**90.8**	**63.7**	52.6	70.1	**76.7**	**96.0**	**95.0**	83.9
Ministral-8B	50.0	45.6	74.5	83.0	55.0	58.6	63.9	68.8	93.0	93.4	84.3
gemma-2-9b	13.4	33.1	3.8	7.4	6.3	35.5	26.7	58.0	43.1	34.2	63.6
SauerkrautLM-9b	60.0	47.5	73.5	88.3	51.7	57.3	56.0	68.8	93.4	92.7	86.2
Teuken-7B	31.0	24.5	57.0	25.6	26.8	24.3	15.8	58.0	73.7	74.5	69.1
SotA	88.4	43.0	62.7	76.0	85.1	65.0	77.3	76.4	92.1	97.7	94.3

Note: The columns represent the datasets, categorized into three parts: humanities datasets, contemporary Twitter datasets and contemporary review datasets. The rows list the tested models, grouped by their associated methods. Comparative SotA values taken from *BBZ* (Borst et al. 2023), *GND* (Zehe et al. 2017), *Lessing* (Schmidt and Burghardt 2018b), *SLK* (Du and Mellmann 2019), *GermEval* (Idrissi-Yaghir et al. 2023), *SB10k* (Barbieri et al. 2022), *Amazon* (Manias et al. 2023) and *PotTs*, *Filmstarts*, *Holidaycheck* and *SCARE* (Guhr et al. 2020). The best value for each dataset is shown in bold.

notable exceptions. Most of the zero-shot approaches outperform the dictionary baseline, and some even reach performance levels of previous SotA scores. However, the exact performance on a given dataset seems to depend not only on the method and task, but also significantly on the choice of the specific model.

Another notable observation is that all zero-shot models and dictionary-based approaches show slightly lower performance on Twitter datasets. This underperformance is likely due to the aforementioned problems associated with the low quality of language in Twitter data. In domains where large amounts of training data are available, such as the 200 million tweets used to train *XLM-T* (Barbieri et al. 2022), or in cases where the data has a high degree of specificity, such as the qualitative *BBZ* dataset, the classification performance achieved by fine-tuned SotA language models remains superior to that of any zero-shot approach. A similar trend is observed in the mixed-domain *GermEval* dataset, given the performance gap between the SotA of 85.1% – achieved using a fine-tuned language model – and our best result with *Llama 70B* at 63.7%, as well as the performance of *XLM-T*. This further supports the notion that specifically fine-tuned models are likely to continue to outperform LLM-based zero-shot approaches on highly specialised tasks (Kheiri and Karimi 2024; Wang et al. 2023; Zhang et al. 2024) Building on previous findings (Borst et al. 2023), it is evident that dictionaries are no longer competitive in terms of classification performance, especially when modern dialogue-based language models are used. Moreover, the addition of *XLM-T*, which is primarily trained on Twitter data, reinforces the observation that the generalisation capability of earlier language models remains limited, even when task-specific refinements are applied to divergent domains.

Many models exhibit performance outliers, and even those that generally perform well tend to have weaknesses on at least one dataset. For example, while NLI-based models typically perform well, they show weaknesses on the *GermEval* and *SB10k* datasets. It is worth highlighting once again the strong performance of NLI models on humanities datasets. In particular, compared to other underperforming zero-shot approaches, such as the Siamese models, the relatively small NLI models are able to compete with dialogue-based models. Within the family of dialogue-based models, *gemma-2-9b* stands out as a significant performance outlier. Of all the variants tested, it has by far the weakest performance, with the model failing to produce valid responses in the majority of cases. The main reason for this poor performance probably is the fact that in our evaluation setup, *gemma-2-9b* is the only monolingual English language LLM that did not receive any specific pre-training on German content. In contrast, *SauerkrautLM-9b*, which originates from the same model family but has been explicitly fine-tuned for the German language, achieved strong results. Although this requires further analysis, it suggests that even in the era of LLMs, language barriers remain a significant factor and cannot be completely ignored. This aspect should not be underestimated, especially with regard to less widely used languages. On the other hand, the *Teuken* model, despite its extensive training in German, cannot compete with privately developed models. All the other LLMs show strong performance, albeit with some fluctuations. For this reason, we have adopted a broader perspective to facilitate the comparison of performance. In order to assess broad applicability, performance in different domains was evaluated relative to the previously established best results (SotA).

When analysing the aggregated performance, as shown in Table 4, certain models stand out as being particularly well suited. While the two largest models achieve the highest classification performance, as expected, the performance gap between them and language models in the range of 8–9 billion parameters is not significant on a global scale. The best overall performer is *Llama-3.1-70B*, which achieves an average of approximately 97% of SotA performance, making it a reliable default choice for many sentiment classification tasks. It is closely followed by the proprietary OpenAI models *gpt-4o* and *gpt-3.5-turbo*, as well as the smaller LLMs *Llama-3.1-8B*, *SauerkrautLM-9b* and *Mistral-8B*. Although the OpenAI models perform very well in comparison, they do not outperform the *Llama-3.1-70B* model, nor do they show a substantial advantage over smaller models, despite being by far the largest, with at least 175 billion parameters. NLI models follow closely, with BERT achieving about 84% of SotA performance. These models, particularly on humanities datasets, perform comparably to results reported in the literature, whereas dictionaries as well as Siamese approaches fall significantly short.

From a purely performance-oriented perspective dialogue-based and NLI methods prove to be robust options for sentiment classification in the German language domain. Humanities research often faces the challenge of working with small but highly domain- and task-specific datasets. In such cases, zero-shot approaches, which do not require fine-tuning or retraining, can offer significant advantages. Beyond performance considerations, however, there are important limitations to consider when evaluating these models and methods. As one moves from efficient and interpretable dictionary-based approaches to large, closed-source language models such as *gpt-4o*, a clear trade-off emerges: increased resource requirements and reduced transparency and interpretability. These trade-offs and their wider implications are discussed in more detail in the following section.

Discussion

In this chapter we critically examine the findings of our evaluation from several perspectives, considering both methodological and practical implications. We first explore the trade-offs between transparency, interpretability and accessibility in different sentiment classification approaches, highlighting the limitations of dictionary-based methods and the challenges posed by the increasing reliance on closed-source LLMs (see Section Transparency, interpretability and accessibility). This is followed by an examination of scalability and resource considerations, where we discuss the computational requirements of different model sizes, the feasibility of running LLMs on consumer-grade hardware, and the implications of quantisation strategies for practical deployment (see Section Scalability and resource considerations). We then consider the impact of non-determinism on inference errors, addressing the variability in LLM outputs, the role of hyperparameter tuning (e.g., temperature settings) and the challenges of ensuring consistency and reliability in zero-shot classification tasks (see Section The impact of non-determinism on model errors). Finally, we discuss the complexities of model selection and performance prediction, assessing how differences in training data, multilingual capabilities and benchmark design affect sentiment classification results, particularly in the context of computational humanities (see Section Model selection and predicting performance). By taking these different perspectives, this discussion aims to provide a nuanced understanding of the strengths, limitations and broader implications of using LLMs for sentiment classification on German-language data.

Transparency, interpretability and accessibility

Dictionaries provide a straightforward, explainable and interpretable approach to sentiment classification while also being highly efficient. Their transparent decision-making processes make them particularly valuable in scenarios where interpretability and low computational cost are required. However, their performance often lags behind more advanced alternatives, especially in nuanced or ambiguous language contexts. While NLI- and dialogue-based approaches provide viable alternatives with superior performance, they also introduce important considerations regarding model interpretability, transparency and accessibility.

Unlike dictionary methods, language model-based classifiers do not provide direct insight into their decision-making processes. Efforts have been made to explain how these classifiers arrive at specific decisions under the umbrella term interpretable AI (Lundberg and Lee 2017; Molnar, 2022; Shrikumar et al. 2017; Simonyan et al. 2014), although these explanations are largely based on indirect mathematical approximations. These methods face additional challenges when applied to text generation models (Amara et al. 2024): At the user level, the ability of generative models to produce not only a classification but also a proposed explanation offers an possibility of tracing the assumed reasoning process. However, this so-called CoT does not accurately reflect the underlying technical processes (Wei et al. 2022). Although this still may be a viable approach for a small number of examples, scaling to larger datasets is likely to result in an overwhelming volume of generated explanations, significantly increasing the effort required for validation and analysis. Beyond performance and interpretability, another layer of consideration is transparency and accessibility. NLI models and non-proprietary LLMs can be used entirely offline, as they are open weight and, in many cases, open-source.[11]

[11] The term *open weight* is used to distinguish between proprietary but downloadable models and truly open source models, which have available source code, report on the data used for training, and have a corresponding license.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Table 4. This table aggregates Table 3, presenting the F1-score averaged across all datasets within each domain

		Humanities	Twitter	Reviews	Avg.
Dictionary	SLK	47.9 (70.9)	47.1 (62.1)	51.6 (57.2)	49.0 (62.8)
	BPW	49.5 (73.3)	48.5 (64.0)	53.1 (58.9)	50.6 (64.8)
	BAWL-R	44.9 (66.4)	34.7 (45.9)	61.4 (68.2)	48.1 (61.7)
	Senti-WS	55.1 (81.6)	40.2 (53.0)	72.5 (80.4)	57.3 (73.6)
	GPC	54.9 (81.2)	41.3 (54.4)	71.3 (79.1)	57.1 (73.2)
Finetuned	germanSentiment	33.4 (49.5)	52.9 (69.7)	80.8 (89.7)	56.0 (71.7)
	XLM-T	46.7 (69.2)	**65.3 (86.1)**	74.525 (82.7)	61.9 (79.3)
NLI	BERT	66.9 (99.0)	39.4 (52.ß)	83.2 (92.2)	65.3 (83.7)
	mDeBERTaX	66.1 (97.9)	37.2 (49.0)	71.1 (78.9)	60.0 (77.0)
Siamese	MPNET-XNLI	59.6 (88.22)	50.1 (66.1)	60.85 (67.5)	57.4 (73.7)
	RoBERTaX	55.6 (82.3)	47.1 (62.0)	61.48 (68.2)	55.4 (71.0)
	RoBERTa-Sentence	54.1 (80.0)	50.3 (66.4)	58.3 (64.7)	54.6 (70)
Dialogue	gpt-3.5-turbo	67.8 (100.0)	57.0 (75.2)	86.1 (95.5)	71.5 (91.7)
	gpt-4o	66.3 (98.1)	63.3 (83.5)	87.8 (97.1)	73.2 (93.9)
	Llama-3.1-8B	62.1 (92.0)	61.5 (81.1)	85.8 (95.2)	70.6 (90.4)
	Llama-3.1-70B	**72.8 (107.8)**	62.1 (81.9)	**87.9 (97.5)**	**75.4 (96.7)**
	Ministral-8B	63.3 (93.7)	59.2 (78.1)	84.9 (94.2)	70.0 (89.8)
	gemma-2-9b	14.4 (21.4)	22.8 (30.1)	49.7 (55.2)	29.6 (37.8)
	SauerkrautLM-9b	67.3 (99.7)	55.0 (72.6)	85.3 (94.1)	70.5 (90.4)
	Teuken-7B	34.5 (51.1)	22.3 (29.4)	68.8 (76.4)	43.7 (56.0)

Note: Additionally, the values in brackets represent the averages obtained after normalising each score by dividing it by the corresponding SotA value for the specific dataset. This normalisation enables a more comprehensive evaluation of the models' performance relative to the established benchmark. The best value for each domain is shown in bold.

In contrast, model transparency is diminished in closed-source systems like *GPT-3.5-turbo* and *GPT-4o*, which are accessible only through *ChatGPT* or the OpenAI API. These proprietary models are primarily commercial products that can only be used according to the manufacturer's specifications and fee structures. While this simplifies accessibility and reduces the need for extensive hardware resources and technical expertise, it lacks direct control and reproducibility, and may even raise concerns about privacy or other ethical issues.

In addition, due to the immense costs associated with training and developing language models, there has been a shift away from the original open-source licensing seen in earlier models such as *BERT* (Devlin et al. 2019) and *DeBERTa* (He et al. 2023), which remain openly accessible, shareable and reusable, thereby improving reproducibility for researchers and practitioners. Although some smaller LLMs can be fine-tuned and have publicly available code, many are no longer fully open-source, but only open weights, due to the specific terms of their individual licences.[12] In general, decreasing openness can lead to many concerns in terms of regulation, legal, or ethical issues (Liesenfeld and Dingemanse 2024; Liesenfeld et al. 2023). In research practice, it remains important to consider this trade-off as part of the selection process: Although proprietary applications such as *ChatGPT* facilitate access to these technologies, the ability to download, view, reuse, share and reproduce them may lead to the choice of an LLM with a higher degree of openness.

Scalability and resource considerations

The size of dialogue-based LLMs varies considerably, ranging from about 1 billion to 405 billion parameters, with *Llama-3.1-70B* being the largest model tested as part of this evaluation study. Running *Llama-70B* at FP16 precision requires approximately 148 GB of VRAM, whereas models with 7–9 billion parameters can be accommodated within approximately 24 GB of VRAM. This implies that smaller models can be run locally for inference on consumer-grade hardware, such as a gaming GPU. In contrast, running the 70B model requires the use of eight enterprise-grade NVIDIA A30 GPUs. Given these hardware constraints, the impact of different quantisation settings in this context warrants further investigation. While 70B models are not easily executable on a single enterprise-grade GPU, even with a reduction in precision to 4-bit quantisation, an 8B model with a modest quantisation to 8 bits can run on a wide range of high-performance consumer and professional devices. Given the performance observed in this study, 8B models such as *Llama-3.1-8B* may represent an optimal balance, offering a trade-off between hardware requirements, performance and open source accessibility.

[12] https://www.llama.com/llama3_1/license/,https://mistral.ai/news/mistral-ai-non-production-license-mnpl/,https://ai.google.dev/gemma/terms

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Table 5. Model runtime and items processed per second on the Amazon dataset

Model	Runtime in s	Items/ s
BERT NLI	102	49.12
Llama-3.1-8B	342	1.62
Llama-3.1-70B	2,737	1.83
Ministral-8B	413	12.11
gemma-2-9b	683	7.27
SauerkrautLM-9b	689	7.26
Teuken-7B	345	14.49

Table 6. Total number of incorrect outputs of all 280,418 data records by model

Model	Total errors	Errors per 1.000 records
Llama-3.1-8B	454	1.62
Llama-3.1-70B	1,122	4.00
Ministral-8B	225	0.80
gemma-2-9b	181,099	654.82
SauerkrautLM-9b	11,349	40.47
Teuken-7B	130	0.46

NLI-based methods, on the other hand, continue to offer a low-resource alternative. Most NLI models are fine-tuned variants of *BERT*, typically containing less than a billion parameters. These models require minimal hardware resources and can even run efficiently on modern laptops. For example, the *BERT* model used in this study consists of only 337 million parameters and fits comfortably within 8 GB of VRAM, which is considered the minimum for today's GPU hardware. Given their strong performance, particularly on humanities datasets, NLI models represent a compelling alternative to larger dialogue-based models. This highlights an important trade-off: the choice between paying directly for access to proprietary systems, such as OpenAI's APIs,[13] versus investing in hardware resources to run models locally. Another important consideration is that inference and, if necessary, training times are highly dependent on the hardware used. While dictionary-based approaches are highly efficient and do not require specialised hardware, neural network-based classifiers benefit greatly from GPU acceleration, often leading to substantial speed improvements. In addition, runtime is affected by the length of the input text, making absolute comparisons difficult. However, while dictionary-based evaluations could be performed in seconds on a standard laptop CPU, the use of language models resulted in considerably longer run times. As shown in Table 5, using the *Amazon dataset* as an example, an increase in the number of parameters corresponds to a noticeable decrease in processing speed. This effect is particularly pronounced for the largest model, *Llama-3.1-70B*, making computational efficiency a crucial factor to consider when using such models.

The impact of non-determinism on model errors

The application of dialogue models to zero-shot text classification has required practical decisions about prompt design and hyperparameter selection. Language models generate text in a non-deterministic way, by sampling the next token from a probability distribution over the vocabulary based on the given input sequence. Consequently, there is an inherent degree of randomness in the generated text, meaning that the same input may produce different outputs upon each execution. This sampling process is controlled by several hyperparameters, of which *temperature* is the most important. Higher temperature values increase randomness, potentially enhancing creativity, while lower values reduce variability, resulting in more deterministic outputs and improved reproducibility.

However, for any given input prompt and text example, the model may produce incorrect outputs that do not match the required structure – in this case, one of the predefined labels. In our study, we found that setting the temperature to zero slightly improved response accuracy, but resulted in a higher error rate. To achieve a balance, we set the temperature to 0.1, aiming to reduce error rates while ensuring that the model remained focused on the task without introducing excessive creativity. An analytical examination of the errors did not reveal any systematic patterns in their occurrence, either across the datasets or across the models, with the notable exception of the *gemma* base model, whose poor performance was primarily due to its high error rate as seen in Table 6.

The absence of German in the training data contributes greatly to the exceptionally high error sensitivity of *gemma*. Even subsequent fine-tuning, as tested with *SauerkrautLM-9b*, leads to significant performance improvements but fails to reduce the error rate to the level of models that were originally trained on multilingual data. Furthermore, the error rate does not always correspond to the classification performance. For example, the *Teuken-7B* model has the lowest error rate but yet delivers the weakest overall results. At the dataset level, the error distribution varies significantly between the models. For example, the best performing model in our test, *Llama-3.1-70B*, produces errors in only 4 out of the 11 datasets tested, but accounts for 99% of the errors in the Holidaycheck dataset. Although a disproportionate number of errors come from datasets with the *neutral* label, there is minimal overlap between the specific datasets that lead to incorrect outputs, suggesting that error patterns are dataset dependent rather than model specific.

Another critical factor influencing both classification performance and error rates is the precise formulation of the prompt. Identifying the optimal prompt is not only a task-specific challenge but also a model-specific optimisation process. As we used a standardised template across all datasets and models in this study, the evaluation revealed considerable variability in the performance of different models. Rather than optimising prompts for specific settings, improvements in prompt design may be better achieved by following established best practices and general guidelines, which remain an active area of research. For sentiment classification, previous experiments with optimised prompts have not produced consistently reliable improvements (Wang and Luo 2023; Wu et al. 2024; Zhang et al. 2024). In addition, techniques that rely on multiple sub-prompts rather than a single prompt – such as some CoT approaches – not only significantly increase runtime but also lead to substantially higher costs when using fee-based services like *GPT-4*. Although certain trends can be inferred from the

[13] The experiments in this study incurred a cost of approximately $77 using the OpenAI Batch API.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

growing body of research, the sheer number of possible parameter variations and the model-specific performance fluctuations make systematic testing almost unmanageable. This challenge is further compounded by the rapid emergence of new model variants, which are introduced on an almost weekly basis.

Model selection and predicting performance

While some models perform well, others underperform significantly, raising the question: what determines model performance? Although larger models tend to perform better in general, there is variability even among models of comparable size. Each model has weaknesses for specific datasets, making it difficult to identify generalisable patterns. The broad applicability of LLMs has led to the development of increasingly complex and combined benchmarks. While these benchmarks provide a useful aggregated comparison of models, they often lack meaningful insights for specific use cases, limiting their practical applicability in individual scenarios. This highlights the ongoing challenge of predicting a model's overall ability to perform specific tasks based on its advertised performance or its pre-training factors.

As with the first transformer models, the initial training of LLMs requires vast amounts of data, yet the details of this data are often not reported transparently. This lack of transparency is particularly problematic for non-English languages. Even in relatively open multilingual models such as *Teuken*, English content remains dominant. For the *Llama* family, the exact composition of training languages is completely unknown. Furthermore, benchmark datasets are often included in the training, making it difficult to ascertain whether predictions on publicly available datasets truly represent zero-shot classification. While this may not be a critical issue for practical applications, it highlights a broader dilemma regarding the transparency and the integrity of model evaluation.

When the generic performance metrics used to promote models do not provide clear insight into their suitability for a specific application, annotated data is required to evaluate classification performance. The mixed results for dialogue-based LLMs – where neither the industry leader *GPT* nor any other model showed clear dominance over its competitors, including NLI models – emphasise the limitations of standard evaluation metrics. These results suggest that commonly used benchmarks do not necessarily indicate the best performing model for sentiment analysis in specific humanities datasets. It also implies that the results obtained for German-language texts may not be directly transferable to other languages.

Conclusion

In this work, we have conducted a comprehensive evaluation of sentiment analysis techniques for German-language data, comparing dictionary-based methods, fine-tuned models and various zero-shot approaches, with a particular focus on dialogue-based LLMs. The results confirm that while fine-tuned models still provide the best performance in data-rich environments, their applicability is limited by the need for extensive training data and computational resources, and the adaptability to other domains is severely restricted. Although newly developed LLMs demonstrate remarkable performance in various contexts, their systematic application introduces a variety of new factors that need be considered.

While proprietary models such as *GPT-4o* perform well, they do not consistently outperform open-weight alternatives such as *Llama-3.1-70B*, which achieved the highest overall classification performance. This suggests that local, open-weight LLMs can serve as a viable alternative to proprietary models, provided that sufficient computational resources are available. The high performance of the models is accompanied by high resource consumption. While the smaller LLMs can still be operated with powerful consumer hardware, the best model we tested can only be used in a high performance cluster environment. However, the computational cost of running large models remains a critical factor, with smaller models such as *Llama-3.1-8B* or *SauerkrautLM-9b* offering a more practical balance between efficiency and performance. In some cases the differences in performance compared to the much less resource-intensive NLI approaches were marginal, suggesting that the number of parameters does not necessarily lead to better results. While the low-threshold nature of text input lowers barriers to entry, it also means that the generalisability of the results is limited by new parameters such as temperature and individual prompts. LLMs present challenges in terms of non-determinism, prompt sensitivity and resource requirements, making their systematic evaluation and deployment difficult. The unpredictability of model performance across different datasets emphasises the limitations of commonly used benchmarks for LLMs in assessing real-world applicability. The high complexity of the models also reduces transparency on several levels. First, the high cost of developing LLMs are increasingly pushing language models into the role of a commercial product whose components are restrictively hidden. Individual fine-tuning or even in-house development for specific research purposes is also almost impossible for individual academic institutions to finance, resulting in a considerable dependence on the published models. Second, the combined benchmarks used in LLM rankings are limited in their ability to indicate aptitude for a particular task or domain. This is even more apparent when the language is not English. Although the multilingual models for sentiment analysis had no problems with German, the English-only model showed that language specificity does exist. Conclusions about other languages are therefore not possible due to the unclear proportions of language composition in the training.

Future research should include further systematic investigation of prompt optimisation, quantification, or the potential of these models under the few-shot paradigm. Nevertheless, this study shows that, with the necessary precautions and appropriate hardware, the use of LLMs for sentiment analysis in German is a promising alternative to both dictionary-based approaches and fine-tuned models.

Data availability statement. The code needed to replicate the experiments and information on where and how to get the data, as well as detailed classification reports, can be found here: https://github.com/JK67FARE/llm_sentiment_german.

Disclosure of use of AI tools. This manuscript was prepared solely by the authors. ChatGPT4o and DeeplWrite were used to improve language and style, and to enhance clarity and readability. However, all original ideas, analyses and original chapter drafts are the genuine work of the authors. Occasionally, ChatGPT4o was used to draft low-novelty text (see the ACL 2023 Policy on AI Writing Assistance https://2023.aclweb.org/blog/ACL-2023-policy/), for instance for the abstract. No AI-generated content was included without human verification and revision. The authors take full responsibility for the final text.

Author contributions. J.K.: Conceptualization, data curation, formal analysis, investigation, software, validation, writing – original draft. J.B.-G.: Conceptualization, data curation, formal analysis, investigation, software, validation, writing – original draft. M.B.: Conceptualization, funding acquisition, supervision, writing – review and editing. All authors approved the final submitted draft.

https://doi.org/10.1017/chr.2025.10005 Published online by Cambridge University Press

Funding statement. This research was enabled through previous funding by the German Research Foundation (DFG) under project number BU 3502/1-1.

Competing interests. The authors declare none.

References

Amara, Kenza, Rita Sevastjanova, and Mennatallah El-Assady. 2024. "Challenges and Opportunities in Text Generation Explainability." In xAI. 2024, Valletta, edited by **Luca Longo, Sebastian Lapuschki, and Christin Seifert**. Cham: Springer Nature Switzerland.

Bannier, Christina, Thomas Pauls, and Andreas Walter. 2019. "Content Analysis of Business Communication: Introducing a German Dictionary." *Journal of Business Economics* 89, no. 1, 79–123.

Bao, Yujia, Menghua Wu, Shiyu Chang, and Regina Barzilay 2020. "Few-Shot Text Classification with Distributional Signatures." In *International Conference on Learning Representations*.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29, no. 1, 19–42.

Barbieri, Francesco, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by **Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis**, 258–66. Marseille: European Language Resources Association.

Borst, Janos, Jannis Klähn, and Manuel Burghardt. 2023. "Death of the Dictionary?—The Rise of Zero-Shot Sentiment Classification." In *Proceedings of the Computational Humanities Research Conference 2023*, Volume 3558 of CEUR Workshop Proceedings, edited by **Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska**, 303–19. Paris: CEUR Workshop Proceedings.

Borst, Janos, Lino Wehrheim, and Manuel Burghardt. 2023. "Money Can't Buy Love?" Creating a Historical Sentiment Index for the Berlin Stock Exchange, 1872–1930." In *Digital Humanities 2023: Book of Abstracts: Zenodo*, edited by **Anne Baillot, Toma Tasovac, Walter Scholger, and Georg Vogeler**, 365–67. Belval/Trier: Zenodo

Borst, Janos, Lino Wehrheim, Andreas Niekler, and Manuel Burghardt. 2023. "An Evaluation of a Zero-Shot Approach to Aspect-Based Sentiment Classification in Historic German Stock Market Reports." In *FedCSIS (Communication Papers)*, 51–60.

Boukes, Mark, Bob van de Velde, Theo Araujo, and Rens Vliegenthart. 2020. "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools." *Communication Methods and Measures* 14, no. 2, 83–104.

Bragg, Jonathan, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. "FLEX: Unifying Evaluation for Few-Shot NLP." In *NeurIPS 2021*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, vol. 33, edited by **H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin**, 1877–901. New York: Curran Associates, Inc.

Campregher Paiva, Isadora, and Josephine Diecke. 2024. "Revisiting Weimar Film Reviewers' Sentiments: Integrating Lexicon-Based Sentiment Analysis With Large Language Models." *Journal of Cultural Analytics* 9, no. 4.

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, edited by **Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens**, 1–14. Vancouver: Association for Computational Linguistics.

Chan, Chung-Hong, Joseph Bajjalieh, Loretta Auvil, Hartmut Wessler, Scott Althaus, Kasper Welbers, Wouter van Atteveldt, and Marc Jungblut. 2021. "Four Best Practices for Measuring News Sentiment Using 'Off-the-Shelf' Dictionaries: A Large-Scale P-Hacking Experiment." *Computational Communication Research* 3, no. 1, 1–27.

Cieliebak, Mark, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. "A Twitter Corpus and Benchmark Resources for German Sen- timent Analysis." In *Proceedings of the Fifth International Workshop On Natural Language Processing for Social Media*, edited by **L.-W. Ku and C.-T. Li**, 45–51. Stroudsburg, PA: Association for Computational Linguistics.

Dennerlein, Katrin, Thomas Schmidt, and Christian Wolff. 2023. "Computational Emotion Classification for Genre Corpora of German Tragedies and Comedies From 17th To Early 19th Century." *Digital Scholarship in the Humanities* 38, no. 4, 1466–81.

Devlin, Jacob, 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)*, edited by **J. Jill Burstein, Christy Doran, and Thamar Solorio**. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.

Dobbrick, Timo, Julia Jakob, Chung-Hong Chan, and Hartmut Wessler. 2022. "Enhancing Theory-Informed Dictionary Approaches With "Glass-Box" Machine Learning: The Case Of Integrative Complexity in Social Media Comments." *Communication Methods and Measures* 16, no. 4, 303–320.

Du, Keli, and Katja Mellmann. 2019. "Sentimentanalyse Als Instrument Literaturgeschichtlicher Rezeptionsforschung." Ein Pilotprojekt. DARIAH-DE Working Papers 32.

Etxaniz, Julen, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. "Do Multilingual Language Models Think Better In English?" In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, edited by **Kevin Duh, Helena Gomez, and Steven Bethard**, 550–64. Mexico City: Association for Computational Linguistics.

Feldkamp, Pascale, Jan Kostkan, Ea Overgaard, Mia Jacobsen, and Yuri Bizzoni. 2024. "Comparing Tools for Sentiment Analysis of Danish Literature From Hymns to Fairy Tales: Low-Resource Language and Domain Challenges." In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, edited by **Orphée de Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi**, 186–99. Stroudsburg, PA: Association for Computational Linguistics.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers For Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120, no. 30, e2305016120.

Guhr, Oliver, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. "Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems." In *Proceedings of the Twelfth Language Resources And Evaluation Conference*, edited by **Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, and Jan Odijk, Stelios Piperidis**, 1627–32. Marseille: European Language Resources Association.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training With Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

Huang, Haoyang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. "Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability By Cross-Lingual-Thought Prompting." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, edited by **Houda Bouamor, Juan Pino, and Kalika Bali**, 12365–94. Singapore: Association for Computational Linguistics.

Idrissi-Yaghir, Ahmad, Henning Schäfer, Nadja Bauer, and Christoph M. Friedrich. 2023. "Domain Adaptation of Transformer-Based Models Using Unlabeled Data for Relevance and Polarity Classification of German Customer Feedback." *SN Computer Science* 4, no. 2, 142.

Jiang, Albert Q., 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."

Jim, Jamin Rahman, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and M. F. Mridha. 2024. "Recent Advancements and Challenges of Nlp-Based Sentiment Analysis: A State-of-the-Art Review." *Natural Language Processing Journal* 6, 100059.

Jin, Renren, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. "A Comprehensive Evaluation of Quantization Strategies For Large Language Models." In *Findings of the Association for Computational Linguistics: ACL 2024*, edited by **Lun-Wei Ku, Andre Martins, and Vivek Srikumar**, 12186–215. Bangkok: Association for Computational Linguistics.

Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. "The Multilingual Amazon Reviews Corpus." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by **Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu**, Online, 4563–8. Association for Computational Linguistics.

Kheiri, Kiana, and Hamid Karimi. 2024. "SentimentGPT: Leveraging GPT for Advancing Sentiment Analysis." In *2024 IEEE International Conference on Big Data (BigData)*, edited by **Wei Ding, Chang-Tien Lu, Fusheng Wang, Liping Di, Kesheng Wu, Jun Huan, Raghu Nambiar, Jundong Li, Filip Ilievski, Ricardo Baeza-Yates, and Xiaohua Hu**, 7051–60. Washington, DC: IEEE.

Kim, Evgeny, and Roman Klinger. 2019. "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies." In Zeitschrift fur digitale Geisteswissenschaften. Erstveroffentlichung vom 16.12.2019. Version 2.0 vom 23.07.2021.Wolfenbuttel 2021. https://doi.org/10.17175/2019_008_v2.

Kocoń, Jan, Igor Cichecki, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. "ChatGPT: Jack of All Trades, Master of None." *Information Fusion* 99, 101861.

Kolb, Thomas, Sekanina Katharina, Bettina M. J. Kern, Julia Neidhardt, Tanja Wissik, and Andreas Baumann. 2022. "The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers." In *Proceedings of the Thirteenth Language Resources And Evaluation Conference (LREC 2022)*, edited by **Nicoletta Calzolari**, 4708–16. Marseille: European Language Resources Association.

Koto, Fajri, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. "Zero-Shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers)*, edited by **Yvette Graham and Matthew Purver**, 298–320. St. Julian's: Association for Computational Linguistics.

Lee, Sanguk, Siyuan Ma, Jingbo Meng, Jie Zhuang, and Tai-Quan Peng. 2022. "Detecting Sentiment Toward Emerging Infectious Diseases on Social Media: A Validity Evaluation Of Dictionary-Based Sentiment Analysis." *International Journal of Environmental Research and Public Health* 19, no. 11, 6759.

Liang, Yixuan, Yuncong Liu, Boyu Zhang, Christina Dan Wang, and Hongyang Yang. 2024. "FinGPT: Enhancing Sentiment-Based Stock Movement Prediction With Dissemination-Aware And Context-Enriched LLMs." Unpublished.

Liesenfeld, Andreas, and Mark Dingemanse. 2024. "Rethinking Open Source Generative AI: Open-Washing And the EU AI Act." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 1774–87. New York, NY: Association for Computing Machinery.

Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse. 2023. "Opening Up ChatGPT: Tracking Openness, Transparency, And Accountability in Instruction-Tuned Text Generators." In *Proceedings of the 5th International Conference On Conversational User Interfaces, CUI '23*. New York, NY: Association for Computing Machinery.

Liu, Chaoqun, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. "Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 9594–9614, Albuquerque, New Mexico.

Liu, Peiyu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024. "Do Emergent Abilities Exist in Quantized Large Language Models: An Empirical Nicoletta Calzolari." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by **Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue**, 5174–90. Torino: ELRA and ICCL.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey Of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55, no. 9, 1–35.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach."

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, vol. 30, edited by **I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett**. New York: Curran Associates, Inc.

Ma, Tingting, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. "Issues With Entailment-Based Zero-Shot Text Classification." In *Proceedings of the 59th Annual Meeting of The Association for Computational Linguistics and The 11th International Joint Conference on Natural Language Processing, (Volume 2: Short Papers)*, edited by **Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli**, 786–96. Association for Computational Linguistics.

Manias, George, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. 2023. "Multilingual Text Categorization and Sentiment Analysis: A Comparative Analysis of the Utilization Of Multilingual Approaches for Classifying Twitter Data." *Neural Computing and Applications*, no. 35, 21415–21431.

McGillivray, Barbara. 2021. "Computational Methods for Semantic Analysis Of Historical Texts." In *Routledge International Handbook of Research Methods in Digital Humanities*, Routledge International Handbooks, edited by **Kristen Schuster and Stuart E. Dunn**, 261–74. London: Routledge Taylor & Francis.

Mengelkamp, Aaron, Kevin Koch, and Matthias Schumann. 2022. "Creating Sentiment Dictionaries: Process Model And Quantitative Study for Credit Risk." In *Proceedings of the 9th European Conference on Social Media*, Number 1, 121–29.

Miah, Md Saef Ullah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and M. F. Mridha. 2024. "A Multimodal Approach to Cross-Lingual Sentiment Analysis With Ensemble of Transformer and LLM." *Scientific Reports* 14, no. 1, 9603.

Molnar, Christoph. 2022. *Interpretable Machine Learning* (2 ed.) Independently published (February 28, 2022). Leanpub.

Mueller, Aaron, and Mark Dredze. 2021. "Fine-Tuning Encoders for Improved Monolingual And Zero-Shot Polylingual Neural Topic Modeling." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by **Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou**, Online, 3054–68. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.

Müller, Karsten. 2022. "German Forecasters' Narratives: How Informative Are German Business Cycle Forecast Reports?" *Empirical Economics* 62, no. 5, 2373–415.

Müller, Thomas, Guillermo Pérez-Torró, Angelo Basile, and Marc Franco-Salvador. 2022. *Active Few-Shot Learning With FASL*, 98–110. Berlin: Springer International Publishing.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and

Ryan Lowe. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Red Hook, NY: Curran Associates Inc. Article 2011, 27730–27744.

Palmer, Matthias, Jan Roeder, and Jan Muntermann. 2022. "Induction of a Sentiment Dictionary for Financial Analyst Communication: A Data-Driven Approach Balancing Machine Learning and Human Intuition." *Journal of Business Analytics* 5, no. 1, 8–28.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, edited by **Jan Hajič, and Yuji Matsumoto**, 79–86. Prague: Association for Computational Linguistics.

Peng, Baolin, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. "GODEL: Large-Scale Pre-Training For Goal-Directed Dialog." arXiv.

Pöferlein, Matthias. 2021. "Sentiment Analysis of German Texts in Finance: Improving and Testing the BPW Dictionary." *Journal of Banking and Financial Economics 2021* 2, no. 16, 5–24.

Přibáň, Pavel, and Josef Steinberger. 2022. "Czech Dataset for Cross-Lingual Subjectivity Classification." In *Proceedings of the Thirteenth Language Resources And Evaluation Conference*, edited by **Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, Stelios Piperidis**, 1381–91. Marseille: European Language Resources Association.

Puschmann, Cornelius, Hevin Karakurt, Carolin Amlinger, Nicola Gess, and Oliver Nachtwey. 2022. "Rpc-Lex: A Dictionary to Measure German Right-Wing Populist Conspiracy Discourse Online." *Convergence (London, England)* 28, no. 4, 1144–71.

Rauchegger, Christoph, Sonja Mei Wang, and Pieter Delobelle. 2024. "OneLove Beyond the Field—A Few-Shot Pipeline For Topic and Sentiment Analysis During the FIFA World Cup in Qatar." In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, edited by **Pedro Henrique Luz de Araujo, Andreas Baumann, Dagmar Gromann, Brigitte Krenn, Benjamin Roth, and Michael Wiegand**, 349–57. Vienna: Association for Computational Linguistics.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-Bert: Sentence Embeddings Using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by **Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan**, 3982–92. Hong Kong: Association for Computational Linguistics.

Reiss, Michael V. 2023. "Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark." arXiv.

Remus, Robert, Uwe Quasthoff, and Gerhard Heyer. 2010. "SentiWS—A Publicly Available German-Language Resource for Sentiment Analysis." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. edited by **Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias**, Valletta: European Language Resources Association (ELRA).

Sarkar, Anindya, Sujeeth Reddy, and Raghu Sesha Iyengar. 2019. "Zero-Shot Multilingual Sentiment Analysis Using Hierarchical Atten- tive Network and BERT." In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, NLPIR 2019*, 49–56. New York, NY: Association for Computing Machinery.

Schmidt, Thomas, and Manuel Burghardt. 2018a. "An Evaluation of Lexicon-Based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing." In *Proceedings of the Second Joint SIGHUM Workshop On Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, August 25, 2018, Santa Fe, New Mexico, USA*, edited by **Beatrice Alex**, 139–49. Stroudsburg, PA: Association for Computational Linguistics.

Schmidt, Thomas, and Manuel Burghardt. 2018b. "An Evaluation of Lexicon-Based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing." In *Proceedings of the Second Joint SIGHUM Workshop On Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, August 25, 2018, Santa Fe, New Mexico, USA*, edited by **Beatrice Alex**, 139–49. Stroudsburg, PA: Association for Computational Linguistics.

Schmidt, Thomas, Manuel Burghardt, and Katrin Dennerlein. 2018. "Kann Man Denn Auch Nicht Lachend Sehr Ernsthaft Sein?– Zum Einsatz Von Sentiment Analyse-Verfahren Für Die Quantitative Untersuchung Von Lessings Dramen." In *Book of Abtracts DHD 2018*.

Schmidt, Thomas, Johanna Dangel, and Christian Wolff. 2021. *SentText: A Tool for Lexicon-Based Sentiment Analysis In Digital Humanities*. ACL: Stroudsburg, Pennsylvania, USA: Universität Regensburg.

Schmidt, Thomas, Katrin Dennerlein, and Christian Wolff. 2021. "Emotion Classification in German Plays With Transformer- Based Language Models Pretrained On Historical and Contemporary Language." In *Proceedings of the 5th Joint SIGHUM Workshop On Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, edited by **Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz**, 67–79. Punta Cana: Association for Computational Linguistics.

Schönfeld, Edgar, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. "Generalized Zero- And Few-Shot Learning via Aligned Variational Autoencoders." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8239–47. Los Alamitos, California: IEEE.

Schwartz, Roy, Jesse Dodge, Noah Smith, and Oren Etzioni. 2019. "Green AI." *Communications of the ACM* 63, 54–63.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. "Learning Important Features Through Propagating Activation Differences." In *Proceedings of the 34th International Conference On Machine Learning—Volume 70, ICML'17*, 3145–53. JMLR.org.

Shu, Lei, Hu Xu, Bing Liu, and Jiahua Chen. 2022. "Zero-Shot Aspect-Based Sentiment Analysis."

Sidarenka, Uladzimir. 2016. "PotTS: The Potsdam Twitter Sentiment Corpus." In *Proceedings of the Tenth International Conference On Language Resources and Evaluation (LREC'16)*, edited by **Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis**, 1133–41. Portorož: European Language Resources Association (ELRA).

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." In *Workshop at International Conference on Learning Representations*.

Socher, Richard, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. 2013. "Zero-Shot Learning Through Cross-Modal Transfer." In *Advances in Neural Information Processing Systems*, vol. 26. New York: Curran Associates, Inc.

Sprugnoli, Rachele, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2016. "Towards Sentiment Analysis for Historical Texts." *Digital Scholarship in the Humanities* 31, no. 4, 762–72.

Stoll, Anke, Lena Wilms, and Marc Ziegele. 2023. "Developing an Incivility Dictionary for German Online Discussions – A Semi-Automated Approach Combining Human and Artificial Knowledge." *Communication Methods and Measures* 17, no. 3, 1–19.

Sänger, Mario, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. "SCARE—The Sentiment Corpus of App Reviews With Fine-Grained Annotations in German." In *Proceedings of the Tenth International Conference On Language Resources and Evaluation (LREC 2016)*, edited by **Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis**. Paris: European Language Resources Association (ELRA).

Taboada, Maite. 2016. "Sentiment Analysis: An Overview From Linguistics." *Annual Review of Linguistics* 2, no. 1, 325–47.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample. 2023. "LLaMA: Open and Efficient Foundation Language Models."

Törnberg, Petter. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning."

van Atteveldt, Wouter, Mariken A. C. G. van der Velden, and Mark Boukes.

2021. “The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, And Machine Learning Algorithms.” *Communication Methods and Measures* 15, no. 2, 121–40.

Veeranna, Sappadla Prateek, Jinseok Nam, Eneldo Loza Mencıa, and Johannes Furnkranz. 2016. “Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents.” *Computational Intelligence* 6, 6–15.

Võ, Melissa L. H. , Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. “The Berlin Affective Word List Reloaded (Bawl-R).” *Behavior Research Methods* 41, no. 2, 534–38.

Waltinger, Ulli. 2010. “GermanPolarityclues: A Lexical Resource for German Sentiment Analysis.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. Electronic Proceedings.

Wang, Yajing, and Zongwei Luo. 2023. “Enhance Multi-Domain Sentiment Analysis of Review Texts Through Prompting Strategies.” In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, 1–7. IEEE.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. “Generalizing From a Few Examples: A Survey On Few-Shot Learning.” *ACM Computing Surveys* 53, no. 3, 1–34.

Wang, Zengzhi, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023. “Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study.” arXiv.

Wehrheim, Lino, Janos Borst, Bernhard Liebl, Manuel Burghardt, and Mark Spoerer. 2023. “More Than a Feeling: Dataset on Media Sentiment Regarding the Berlin Stock Exchange.”

Wei, Jason, 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.” ArXiv abs/2109.01652.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” In *Proceedings of the 36th International Conference On Neural Information Processing Systems, NIPS ’22*, edited by **S. Koyejo, eS. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh**, Red Hook, NY: Curran Associates Inc.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.” In *Proceedings of the 30th Conference on Pattern Languages of Programs (PLoP ‘23)*. The Hillside Group. Article 5, 1–31.

Widmann, Tobias, and Maximilian Wich. 2022. “Creating and Comparing Dictionary, Word Embedding, And Transformer-Based Models to Measure Discrete Emotions in German Political Text.” *Political Analysis* 31, no. 4, 1–16.

Wojatzki, Michael, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. “GermEval 2017: Shared Task on Aspect- Based Sentiment In Social Media Customer Feedback.” In *Proceedings of the GermEval 2017—Shared Task On Aspect- Based Sentiment in Social Media Customer Feedback*, edited by **Michael Maximilian Wojatzki (Hrsg.), Eugen Ruppert (Hrsg.), Torsten Zesch (Hrsg.), and Chris Biemann (Hrsg.)**, 1–12. Berlin: DuEPublico Duisburg-Essen.

Wolf, Thomas, 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, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. “Transformers: State-of-the-Art Natural Language Processing.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by **Qun Liu and David Schlangen**, 38–45. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.

Wu, Chengyan, Bolei Ma, Zheyu Zhang, Ningyuan Deng, Yanqing He, and Yun Xue. 2024. “Evaluating Zero-Shot Multilingual Aspect-Based Sentiment Analysis With Large Language Models.” arXiv.’

Xian, Yongqin, Bernt Schiele, and Zeynep Akata. 2017. “Zero-Shot Learning - The Good, the Bad and the Ugly.” In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 3077–86.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. “XLNet: Generalized Autoregressive Pretraining For Language Understanding.” In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, edited by **Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alche-Buc, and Emily B. Fox**, 5754–64. Red Hook, NY: Curran Associates Inc.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. “Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, edited by **Jing Jiang, Vincent Ng, and Xiaojun Wan**, 3912–21. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.

Zehe, Albin, Martin Becker, Fotis Jannidis, and Andreas Hotho. 2017. “Towards Sentiment Analysis on German Literature.” In *KI 2017: Advances in Artificial Intelligence*, volume 10505 of Lecture Notes in Computer Science, edited by **Gabriele Kern-Isberner, Johannes Fürnkranz, and Matthias Thimm**, 387–94. Cham, Switzerland: Springer International Publishing.

Zhang, Ranran Haoran, Aysa Xuemo Fan, and Rui Zhang. 2023. “ConEntail: An Entailment-Based Framework For Universal Zero and Few Shot Classification With Supervised Contrastive Pretraining.” In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, edited by **Andreas Vlachos, and Isabelle Augenstein**, 1941–53. Dubrovnik, Croatia: Association for Computational Linguistics.

Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. “Sentiment Analysis in the Era of Large Language Models: A Reality Check.” In *Findings of the Association for Computational Linguistics: NAACL 2024*, edited by **Kevin Duh, Helena Gomez, and Steven Bethard**, 3881–906. Stroudsburg, PA: Association for Computational Linguistics.

Zhang, Xiang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. “Don‘t Trust ChatGPT When Your Question Is Not In English: A Study of Multilingual Abilities And Types of LLMs.” In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by **Houda Bouamor, Juan Pino, and Kalika Bali**, 7915–27. Singapore: Association for Computational Linguistics.

Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. “DIALOGPT: Large-Scale Generative Pre-Training For Conversational Response Generation.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, edited by **Asli Celikyilmaz and Tsung-Hsien Wen**, 270–8. Stroudsburg, PA: Association for Computational Linguistics.

Zhu, Xiliang, Shayna Gardiner, Tere Roldán, and David Rossouw. 2024. “The Model Arena for Cross-Lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models.” In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, edited by **Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi**, 141–52. Bangkok: Association for Computational Linguistics.