Zero-shot prompt-based classification: topic labeling in times of foundation models in German Tweets

Simon Münker **Kai Kugler** **Achim Rettinger**
Trier University, Germany
{muenker, kuglerk, rettinger}@uni-trier.de

Abstract

Filtering and annotating textual data are routine tasks in many areas, including social media and news analytics. Automating these tasks enables scaling analyses with respect to speed and breadth while reducing manual effort. Recent advancements in Natural Language Processing, particularly the success of large foundation models, provide new tools for automating annotation processes through text-to-text interfaces with written guidelines, eliminating the need for training samples.

This work assesses these advancements in a real-world setting by empirically testing them on German Twitter data about social and political European crises. We compare prompt-based results with human annotations and established classification approaches, including Naive Bayes and BERT-based fine-tuning with domain adaptation. Despite hardware limitations during model selection, our prompt-based approach achieves comparable performance to fine-tuned BERT without requiring annotated training data. These findings highlight the ongoing paradigm shift in NLP toward task unification and the elimination of pre-labeled training data requirements.

1 Introduction

Since ChatGPT's release in November 2022, both public and scientific interest has shifted toward generative NLP technologies like Large Language Models (LLMs) (Kalla et al., 2023). Key questions focus on human-machine interaction, specifically the benefits these tools offer for automating manual tasks. Generative foundation models function as multilingual chatbots (Ouyang et al., 2022), following natural language instructions while interpreting texts by statistically capturing human knowledge and replicating language understanding capabilities.

The formulation of these commands, termed "prompt engineering", combined with powerful models, enables solving tasks the model has not been extensively trained on—a capability known as zero- or few-shot learning (Brown et al., 2020). When instruction-following, natural language understanding, and few-shot learning are combined, they promise to significantly reduce manual effort in automating textual data annotation processes.

Unlike traditional supervised learning approaches that require labeled datasets, prompt-based methods leverage the model's general language understanding capabilities through task-specific instructions (Liu et al., 2023). This paradigm shift is particularly relevant given recent research comparing in-context learning and fine-tuning strategies (Min et al., 2022), which demonstrates that language models can achieve competitive performance without task-specific training data.

The approach aligns well with researchers investigating current topics in online social networks. As societal crises increase in frequency (Guterres and Secretary-General, 2022), timely analysis becomes crucial for understanding public opinion tipping points. Projects like SOSEC[1] consult survey participants weekly to track developments, but even weekly updates may miss influential events. LLMs potentially offer a complementary tool matching the temporal and quantitative scale needed for high-frequency analytics.

This work investigates using open pre-trained generative language models to process social media text datasets in real-world conditions. The requested annotations prove challenging even for human annotators despite extensive instructions. Our focus is not building superior annotation approaches regarding overall accuracy, but evaluating how well current LLMs serve as automated primary annotation tools without examples, assuming

[1] SOSEC Project Homepage (last retrieved Jun. 23, 2025): https://www.socialsentiment.org

an experimental setup requiring open local models for control and reproducibility with moderate hardware requirements.

Accordingly, we address the following research questions:

RQ$_1$ Can zero-shot prompt-based classification achieve comparable results to a fine-tuned classifier and align well to human annotations?

RQ$_2$ How does the scope of information provided to the model, i.e. the extent of annotation guideline impact the performance?

In in addition to answering our research questions. We provide a standalone Python module for prompt-based classification with local LLMs (see Sec. 4.3).

2 Background

The motivation for our work is twofold. Content-wise, the political and social situation in the EU poses a relevant interdisciplinary subject. In particular, how citizens express their opinions on online social media platforms. For the scope of this work, we omit a detailed description. Collecting large amounts of unlabeled data comes with the need for annotation to enable future analysis. Streamlining the annotation displays our technical motivation. With the advent of LLMs capable of performing various tasks, new approaches emerged to classify textual data. Notably, methods allow classifying content through a text-2-text interface, where the user can align the classification expectations based on textual annotation guidelines (Brown et al., 2020). That omits the need for machine-learning-based optimization and shifts the focus to formulate human-readable guidelines that the model can follow.

Text classification, like sentiment analysis or topic labeling, holds significant importance in both research and the economy (Petersen-Frey et al., 2023). It enables us to extract valuable insights from textual data and make informed decisions across various domains, including customer feedback analysis, market research, and automated content moderation (Minaee et al., 2021). Traditionally, text classification relied on supervised learning approaches utilizing task specific models (Kadhim, 2019) or fine-tuning a pre-trained models on a labeled datasets (Weißenbacher and Kruschwitz, 2023). The development of optimized and robust text classifiers is therefore a resource-intensive task. Preceding research shows that data-driven classification approaches (Edwards and Camacho-Collados, 2024) outperform prompt-based approaches on a selection of datasets. However, the approach does not provide tailored prompts or incorporate annotation guidelines. In contrast, we focus on a single dataset and conduct a more detailed experiment.

Instruction Fine-tuning The success of LLMs was followed by a paradigm shift triggered by a proposal from Google in 2020 (Raffel et al., 2020a), (Sun et al., 2022). To this point, the typical pipeline combines fine-tuned models like BERT (Vaswani et al., 2017) or XLNet (Yang et al., 2019) with a task-specific classification head. For classification tasks, the attached head architecture produced a probability distribution over the given classes (Kant et al., 2018). For generative tasks, a sequential decoder was used as an attached head, which generates a text sequence as output (Jiang et al., 2021). In contrast, the unified pipeline has three main advantages: a) the optimization pipeline, including the data preparation, is more efficient as the models achieve state-of-the-art performance with less labeled data, b) the approach strengthens the capability of transferring knowledge to unseen tasks using a known formulations, and c) from the non ML researchers perspective, unified models are easier to infer and deploy.

Prompt Engineering Instruction-based model solve tasks that are provided in human-like text during conversations. However, the effectiveness of these models relies heavily on the quality and specificity of prompts given to them. Prompt engineering, the process of formulating and refining prompts, plays a crucial role in harnessing the full potential of LLMs (Liu et al., 2023). Unlike the traditional pipeline for supervised tasks, which trains a model to take in a textual input and predict an output, prompt-based approaches utilize LLMs in a dialog.

This paradigm shift allows us to bypass the aforementioned bottlenecks. We no longer require pre-labeled datasets for fine-tuning the models specifically for each application. Instead, we can utilize the model's general language understanding capabilities and prompt it with task-specific instructions. This significantly reduces the need for large-scale labeled datasets (Sun et al., 2022), which can be

expensive and time-consuming to create.

2.1 Multilingual Considerations and Real-world Challenges

The application of LLMs to non-English content presents additional complexities that are particularly relevant to our work. While many instruction-tuned models are trained on multilingual corpora, their instruction-following capabilities are often predominantly developed using English examples (Muennighoff et al., 2023a). This creates a potential mismatch between the model's general language understanding in various languages and its ability to follow task-specific instructions in those languages.

Furthermore, real-world text classification scenarios often involve noisy, informal, and contextually dependent content—characteristics that are particularly pronounced in social media data. Traditional benchmark datasets may not adequately reflect these challenges, potentially overestimating the performance of both traditional and prompt-based approaches when deployed in practical applications (Bender et al., 2021). Our focus on German Twitter data about political crises represents an attempt to address this gap by evaluating methods under more realistic conditions.

The intersection of multilingual capabilities, instruction following, and real-world data complexity forms the technical foundation for our investigation into zero-shot prompt-based classification as a practical alternative to traditional supervised learning approaches.

3 Data

To assess the capabilities of zero-shot prompt-based classification in a real-world setting, we deliberately did not resort to an academic benchmark, since they tend to not reflect the challenges of real-world topic labeling appropriately. Also, we intended to avoid a standard but unrealistic setting with English only data.

3.1 Collecting

We collected a German Twitter data set according to a topical selection defined by the survey questions of the SOSEC project about the energy crises in the winter of 2022/2023. The non-English data set was picked to further stress-test the LLMs' capabilities in a realistic setup. At that time, Twitter (now 𝕏) still provided API access. We compiled a comprehensive list of hashtags and keywords that broadly reflected the described crises. The list consisted of relevant terms, including trending keywords, hashtag-based identifiers of political parties, and persons of interest. We queried for each keyword in the list consistently between October 2022 and May 2023. During this time, we collected approximately 750,000 samples.

3.2 Manual Annotation

Two domain experts and native speaker annotated a random selection of approx. 7000 tweets. The annotators were instructed accordingly and given a manual with guiding questions on whether a tweet should be annotated or not. Of the selection samples, only 3000 could be annotated as belonging to a topic, as many tweets did not match our criteria. A high degree of noise due to ambiguity, variation, and uncertainty is a common property of real-world data sets (Beck et al., 2020).

4 Methods

The candidate methods we picked for automating the annotation task, are taken from three eras of modern NLP: A Naive Bayes classifier, representing the pre-deep learning era, is picked as the baseline. Next, for the deep learning era, a pre-training and fine-tuning approach using a BERT transformer (Kenton and Toutanova, 2019) is selected. Finally, for the era of foundation models, we use instruction-tuned models based on the transformer T5 (Raffel et al., 2020b). Again, we tried to setup a realistic "in-the-wild" scenario by picking freely available models, that can be run on moderate hardware requirements.

4.1 Baseline

In order to establish a baseline for the methods and our prompt-based classification task, we employ a Multinomial Naive Bayes Classifier (Manning, 2009). To represent our text data numerically, we utilize a count vectorizer also provided by scikit-learn (Pedregosa et al., 2011). The count vectorizer converts the textual data into a matrix of token counts, where each row represents a sample, and each column represents a unique word or token in the corpus.

4.2 Fine-tuned Transformer

We chose the model "gbert-base", for German BERT, which is a language model specifically designed for text classification and Named-entity

recognition in German (Chan et al., 2020). For our tasks, we fine-tune all parameters on 80% of the annotated data as a single class classification task. Upon completion of the model development and training, we deployed the models to the Hugging Face model hub. The models are available under the "anonymized during review" account, allowing other users to access and utilize them for their own applications.

4.2.1 Additional Domain Adaption

To further improve the performance of our fine-tuned classification model, we utilize our raw data (approx. 750,000 tweets). Thus, we include an additional pre-training phase to shift the model's language understanding toward the target domain (Ramponi and Plank, 2020). We shift the focus of the generalized pre-trained BERT model to a Twitter-specific language. That improves the robustness of the model to achieve out-of-distribution generalization without training a model from scratch for our task. The inclusion of a second pre-training phase (adaptive pre-training) improves performance and generation significantly for classification tasks (Manjavacas and Fonteyn, 2022).

4.3 Zero-Shot Prompting

The two preceding methods set the traditional machine-learning baseline and current SOTA for text classification. Our text-to-text zero-shot prompting (Kojima et al., 2022) approach differs in two main aspects. It benefits from the text input and text output paradigm and, thus, pulls away from mathematical optimization. Thereby, we can study the impact of textual formulation on our annotation goal, align the annotation by words, and not optimize by parameters. It does not rely on training data or examples (zero-shot) and, thus, cannot overfit the provided data or assimilate the included biases.

We restrict our setup and the model selection to a level that modern desktop workstations (approx. 5.000€ in 2023) can effectively run the program. With this, we underline the applicability during active research for smaller groups or individuals. For our experiments, we compare a monolingual and a multilingual instruction-tuned model in four different sizes. Regarding the prompts, we analyze the performance of levels of textual detail, from vague introductions to a reduced version of the annotation guidelines.

Model Selection To allow for a reproducible experimental setup we limit our selection to freely available models from the platform Hugging Face supporting English and German and trained in an instruction-tuned text-to-text scenario. With this filter, the selection is reduced – selection date: Mai 2023 – to two models, namely Flan-T5 (Chung et al., 2024) and mT0 (Muennighoff et al., 2023b). Both models are based on the same fine-tuned transformer T5, each fine-tuned and adapted in a custom manner. This selection allows for a comparison and evaluation of the adaption quality beyond prompt templates alone. Both models are available in four different sizes, usable with our restrictions. Thereby, we can compare, in addition, the respective performance across several parameters. It gives us a third dimension of analysis.

Prompts We provide a baseline prompt (Prompt 1) that is generic without a specific task description. The terms in curly braces represent variables, substituted during prompting. To differentiate the task description from the text content, we use triple back-ticks (```) as delimiters (White et al., 2023). Additionally, the template emphasizes choosing a single class through the keywords "categorize" and "one of".

```
prompt: str = f"""
Categorize the following tweet into one
of the listed classes {classes}:
'''{text}'''
"""

classes: List[str]
text: str
```

Prompt 1: base

In the preceding prompt, we omit a naming type of classification task. In the following prompts, we gradually add levels of information. To analyze if and how the models benefit from an additional explanation. In the first prompts, we introduce the name of the respective tasks (Prompt 2). As both models are fine-tuned for various classification tasks, we assume that they benefit from the task names.

In the following two prompts, we give a short description about the task. In addition to naming the task explicitly, we provide additional synonyms for task (Prompt 3.

The last prompt we tested contain a condensed version of the annotation handbook (Prompt 4. We

```
prompt: str = f"""
Your task is to classify the following
tweet regarding its topic into  one of
the following classes {classes}:
'''{text}'''
"""
```

Prompt 2: task-name

```
prompt: str = f"""
Your task is to analyze the topic of the
following tweet:
'''{text}'''
Thus, identify the dominant subject of
the tweet content and classify it into
one of the following classes: {classes}
"""
```

Prompt 3: description

could not use the full version as our models are restricted in the input length, and the complete topic task description would not leave room for the input of the tweet. With this information, we provide the model with nearly identical instructions as the human annotators.

```
prompt: str = f"""
Your task is to utilize the following
class descriptions - label between *'s
followed by its definition - to choose
the one most fitting for the tweet:
*Wirtschaft*: The tweet contains
concerns about the economic crisis or
the personal financial situation.
*Migration*: The tweet evaluates the
chances and dangers of migration and
makes judgmental remarks about migrants
or as migrants perceived people.
*Demokratie*: The tweet expresses trust
or distrust towards the parliament and
advocates or rejects the democratic
system.
*Ukraineunterstützung*: The tweet states
the author's position on the
Russo-Ukrainian war, evaluates economic
penalties against Russia, or postulates
financial or military support for
Ukraine.
*Energiewende*: The tweet discuss
personal concerns about the power supply
or energy system transformation.
'''{text}'''
"""
```

Prompt 4: handbook

Metric In traditional machine learning classification pipelines the model response represents one of the given classes or numerical representation. However, in our prompt-based approach, the models respond with unrestricted free-form text. Thus, the model is not limited to responding with one of the targets but may produce additional explanations or invent new classes. This fact prevents us from utilizing traditional metrics relying on confusion matrices. In our approach, we are not guaranteed to receive a miss classification with a false positive label. We cannot apply metrics relying on type I (false positive) and type II (false negative) errors. Therefore, we restrict our evaluation to the calculation of the macro average (unweighted mean). As we receive a free-form text as a response, we apply further pre-processing to extract the predicted label. We count only exact case-insensitive matches. We exclude responses containing additional characters or leading/trailing spaces.

Implementation We implemented our approach utilizing Hugging Face (Wolf et al., 2019) for model loading and prediction, and handled data flow and results analysis with Pandas (Wes McKinney, 2010). We emphasize that the project is structured to be easily expandable for further LLMs and API integration. We publish our pipeline as a pip repository[2]. The pipeline configuration assumes two main inputs: a list of prompts and a list of models to compare. Each model is queried with each prompt, resulting in multiple experiments. This approach allows for a comprehensive comparison of model performance across different prompts. The querying is performed batch-wise to facilitate efficient and streamlined interactions with the models during the experimentation process. After the querying process, the pipeline uses an automated system for collecting results for each prompt and model combination in every experiment to ensure consistent and reliable data collection. We also include a basic plotting functionality, which assumes a sequential relationship between the two dimensions being analyzed.

5 Results

We utilize local resources to run all experiments. All calculations are performed on a single NVIDIA Tesla V100 32GB GPU combined with two Intel Xeon Silver 12 core 2.2GHZ CPUs and 512GB RAM. We developed our experimental environment to run the predictions batch-wise, looping for every model over every prompt.

[2]Package available on PyPi: https://pypi.org/project/cltrier_promptClassify/

5.1 Methods Comparison

The comparison between the baseline and fine-tuned transformer models reveals a substantial disparity in their classification performance. While the baseline model achieves an approximate weighted average F1 score of 66%, the fine-tuned transformer model achieves approximately 86%, representing a significant difference of 20% (cmp. Figure 1). This contrast emphasizes that the topic possesses an underlying semantic meaning that cannot be effectively captured using a simplistic count-based approach. Instead, the intricate language comprehension capabilities of a transformer model are required to accurately grasp the nuances and subtleties of our topics.

Additionally, we observe variations in performance across different classes for both approaches (cmp. Table 1). Both models exhibit lower performance in classifying tweets related to "Demokratie" (approximately 56% for baseline vs. 77% for fine-tuned transformer) and "Wirtschaft" (approximately 48% for baseline vs. 78% for fine-tuned transformer). In contrast, classes with high F1 scores such as "Energiewende" (approximately 78% for baseline vs. 85% for fine-tuned transformer) and "Ukraineunterstützung" (approximately 75% for baseline vs. 91% for fine-tuned transformer) demonstrate superior classification accuracy. We hypothesize that the topics with higher F1 scores possess more distinct and well-defined terminology, making the classification task easier, particularly for the baseline model.

5.2 Prompting Detail

Our results show that, with more information, the performance gradually improves with the larger models (cmp. Table 2). However, the smaller versions of each family does not profit from the additional information as they struggle to understand the task description in general, and their responses show that the additional information confuses the model and diffuses the given task. Interestingly, solely mentioning the task name noticeably improves the performance compared to the base prompt. We assume that the information is sufficient for the model to connect inside its internal parametric task memory to a similar task from its own instruction-tuning stage. This provides a glimpse into how zero-shot and in-context learning works within foundation models.

6 Discussion

While our classification results show comparable performance to the baselines, we observe new challenges unseen in classic machine-learning pipelines. These represent the typical pitfalls of LLMs.

Hallucinations Independently from the sizes both model families fabricate topics not given in our prompts. In particular, the small and base models suffer from this behavior. We place this phenomenon under the term LLM hallucination. In general, it describes the generation of false information when an LLM has no internal information about a task or question asked. Interestingly, the terminology concerning language models and behavior is criticized, and researchers propose the usage of the word confabulation (Chalmers, 2023). It describes, in a psychiatric context, the behavior of people to invent plausible-sounding justifications that have no basis. These individuals appear to strongly believe in the story and do not intend to deceive with the information (Moscovitch, 1995). The change in terminology and perspective allows for an analysis of the phenomenon in contrast to human behavior and comparison with neural pathologies: "What are LLMs but humans with extreme amnesia and no central coherence?" (Millidge, 2023)

Inconsistencies Our results show a highly inconsistent behavior not only between prompt variations but also for different samples and the same prompt. As we described in our results, the models generate responses that do not match our task description, like translation and code snippets for some prompt templates. However, we observe also the occurrence of these phenomena for individual samples while using prompts that provide mostly sound responses. These inconsistencies occur for both models in all sizes, even though mT0 is more affected. Current research investigates negated prompts and shows that models perform significantly worse (Jang et al., 2023). These results question the task understanding of LLMs and underline how sensitive they are to their inputs. Transferred to our approach, the inconsistencies may be caused by linguistic phenomena inside the Tweets which alters the prompt meaning for the model.

Blackbox With prompt-based approaches, we overall move more in a direction, where the machine learning black box becomes even more

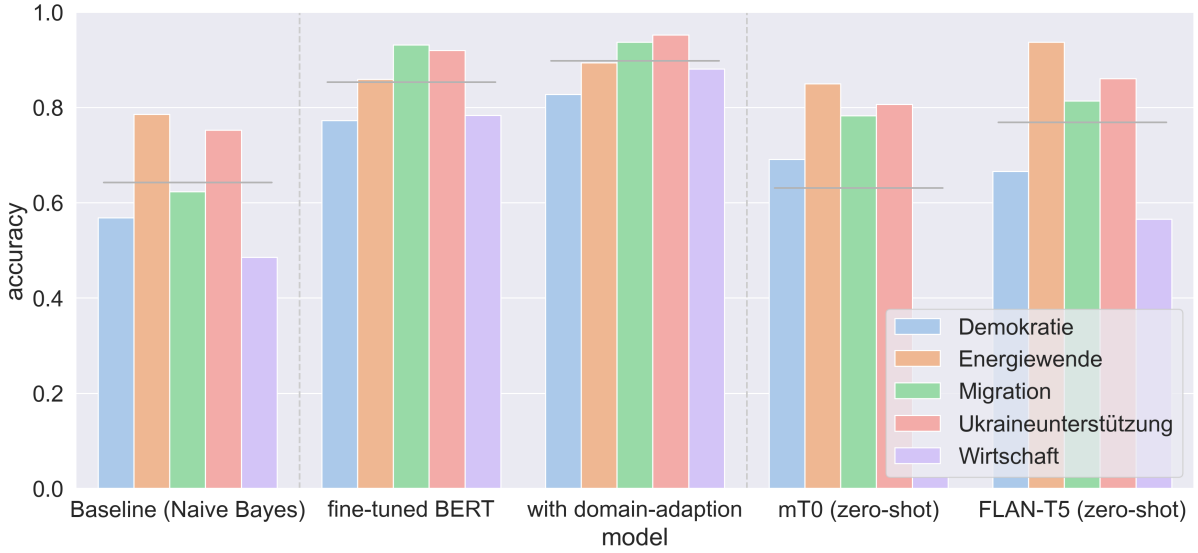

Figure 1: Comparison of different classification methods, showing the accuracy across five political topics, comparing the baseline with a fine-tuned and domain-adapted BERT and two instruction models with zero-shot approaches. The gray lines show the average performance across all classes for a model.

	Baseline Naive Bayes	fine-tuned	**BERT** w/ pre-training	**mT0** zero-shot	**FLAN-T5** zero-shot
Demokratie	0.5684	0.7727	**0.8276**	0.6908	0.6660
Energiewende	0.7857	0.8593	0.8939	0.8500	**0.9368**
Migration	0.6230	0.9310	**0.9367**	0.7826	0.8140
UA-Unterst.	0.7521	0.9199	**0.9524**	0.8066	0.8604
Wirtschaft	0.4857	0.7831	**0.8807**	0.0254	0.5657
macro avg	0.6430	0.8532	**0.8983**	0.6311	0.7686

Table 1: Comparison of different classification methods, showing the accuracy and the macro average comparing the baseline with a fine-tuned and domain-adapted BERT and two instruction models with zero-shot approaches. Highlighted **bold** the best-performing model for each class.

opaque in contrast to traditional ML methods (Ollion et al., 2024), as we cannot see the prediction scores for each possible class. This is a major disadvantage as optimizing the pipeline relies on comparing the textual results with the provided prompts. Combined with the issue that traditional metrics, which rely on the confusion matrices, are inapplicable, a qualitative analysis during the prompt optimization becomes necessary.

Inherent Model Biases LLMs inherit biases present in their training data, which predominantly consists of web-scraped content reflecting societal biases and prejudices (Gallegos et al., 2024). In the context of political and social crisis analysis, as examined in our study, these biases can significantly skew annotation outcomes. For instance, models may exhibit systematic preferences toward certain political viewpoints, demographic groups, or cultural perspectives that were overrepresented in their training data. This is particularly concerning when analyzing German Twitter data about European crises, where models trained predominantly on English content may not adequately capture cultural nuances or may impose Anglo-centric interpretations on German political discourse.

7 Conclusion

Concerning RQ_1, our results show that with a well-defined prompt, including a summarized annotation handbook, our prompt-based approach achieves nearly on-par performance with the fine-tuned baseline and surpasses the naive baseline. When taking into account, that we tested a challenging non-English task in a real-world setting with restrictions in model and context window size, and the early development stage of freely available instruction-based models, we assume that our results will significantly tilt towards LLMs in the future. Thus,

	base		w/ task-name		w/ description		w/ handbook	
	FLAN-T5	**mT0**	**FLAN-T5**	**mT0**	**FLAN-T5**	**mT0**	**FLAN-T5**	**mT0**
Demokratie	0.4389	**0.7595**	0.4389	0.6832	0.5229	0.6908	0.6660	0.0324
Energiewende	0.8559	0.8015	0.8750	0.8206	0.8588	0.8500	**0.9368**	0.6868
Migration	0.9179	0.5990	**0.9203**	0.7150	0.8865	0.7826	0.8140	0.2126
UA-Unterst.	0.7659	0.7000	0.8000	0.7231	0.7330	0.8066	**0.8604**	0.6714
Wirtschaft	0.4640	0.0000	0.4831	0.0064	**0.6017**	0.0254	0.5657	0.1292
macro avg	0.6885	0.5720	0.7035	0.5896	0.7206	0.6311	**0.7686**	0.3465

Table 2: Impact of prompt engineering on zero-shot classification performance, comparing two instruction models across four prompt variants on class-based accuracy and the macro average. The complexity of the prompt increases from left to right. Highlighted **bold** the best-performing model for each class.

we expect that prompt-based text classification will be highly relevant for future use in academia when empirical studies on large quantities of text are conducted.

Concerning RQ_2, analyzing our prompts in detail along the predefined dimension, we found the following: The difference in German and English prompts in the smaller models is especially significant. Only the XL version does understand the German task formulation. Thus, we assume multilingual knowledge is reduced significantly during the parameter pruning. Also, we conclude that instruction training on mostly English tasks does not lead to multilingual task generalization despite pretraining the model on multilingual corpora. Despite not understanding the German task description, the models handled German tweets and classes without any issues. That highlights the importance of the prompt formulation and its closeness to tasks seen during the fine-tuning process.

Manipulating the order of the prompt segments shows only a minor impact on the performance. Inserting the full Tweet into the center of the prompt reduces the quality of the results, which highlights the importance of handling long-distance dependencies. Further, the separation between task and content led to confusion due to the usage of symbols possibly resembling programming code.

Concerning the scope of detail, our results show a correlation between the performance and the extent of information provided in the task description. Larger models benefit more from the detailed description. That aligns with current research on the formulation of prompts and model selection for enhancing the quality of prompt-based tasks (White et al., 2023; Logan IV et al., 2022). In summary, our results support the current techniques for zero-shot prompting proposed in research (Liu et al., 2023) and online learning guides (DAIR.AI, 2023).

7.1 Future Work

Our experiments display the SOTA of Mid 2023. The research around LLMs relevant to our approach expands in two dimensions. On a daily base, new models are released larger in size and higher in performance. We highly recommend extending the research to the recent and more capable LLMs to harness the full potential of prompt-based annotation. The usage of larger models would not only increase the zero-shot performance but also allow more complex prompt variants (Almazrouei et al., 2023), (Touvron et al., 2023). We suggest including examples (few-shot) in prompts to improve the results. We expect a reduction of inconsistencies and hallucinations (Logan IV et al., 2022), coupled with a higher alignment to the annotation intents.

While considering the annotation task in a real-world setting, it also delivers inconsistencies like human annotations, capturing personal and demographic properties of the annotators might lead to a more insightful annotation outcome. This can be achieved by adding personas to the prompt or conditioning LLMs on individual human behavior. Considering the domain of prompt engineering, the proposal adapts the idea of role prompting, which shapes the output style of the generated text resembling a certain person. This adaptation method significantly enhances the quality and accuracy of generated content (White et al., 2023), (Shanahan et al., 2023).

In summary, the potential for mimicking human behavior in text annotation tasks with LLMs seems enormous. While providing computational social science researchers with a powerful new tool, it also opens up many critical uses like personalized opinion manipulation and impersonation. Potentials for abuse have to be closely monitored.

Limitations

Our study acknowledges several important limitations that constrain the generalizability and applicability of our findings:

Language and Cultural Specificity: While we intentionally chose German Twitter data to stress-test multilingual capabilities, our findings may not generalize to other languages or cultural contexts. The models' performance on German content, particularly with smaller model sizes, revealed significant limitations in multilingual task understanding that may vary across different language pairs and cultural domains

Temporal Constraints: Our data collection period (October 2022 to May 2023) represents a specific temporal snapshot of political and social discourse. The topics and language patterns during the European energy crisis may not reflect classification challenges in other time periods or crisis contexts, limiting the temporal generalizability of our approach.

Annotation Subjectivity: Despite providing extensive annotation guidelines, the inherent subjectivity in topic classification tasks, particularly for political and social content, introduces variability that affects both human baseline annotations and model evaluation. The high degree of noise in real-world social media data, with only 3,000 out of 7,000 initially selected tweets meeting annotation criteria, highlights the challenging nature of the task.

Evaluation Methodology: Our restriction to exact case-insensitive matches for model outputs, while necessary given the free-form nature of LLM responses, may have been overly conservative and potentially underestimated model performance. The inability to apply traditional confusion matrix-based metrics limits our ability to conduct nuanced error analysis.

Ethical Considerations

Our research raises several ethical considerations that warrant careful attention. LLMs inherit and potentially amplify biases present in their training data, which predominantly consists of web-scraped content reflecting existing societal prejudices. In our context of analyzing German political discourse about European crises, these models may systematically favor certain political viewpoints, demographic perspectives, or cultural interpretations that were overrepresented during training. This bias propagation is particularly concerning when models trained primarily on English content are applied to German political discourse, potentially imposing Anglo-centric interpretations on European political contexts.

The automation of political and social content classification also raises fundamental questions about the appropriate role of AI systems in interpreting politically sensitive discourse. It may inadvertently contribute to the depersonalization of political analysis and reduce human oversight in contexts where nuanced cultural and political understanding is crucial. This concern extends to the "black box" nature of LLMs, which creates challenges for accountability in automated annotation decisions. Unlike traditional machine learning approaches where prediction scores provide some interpretability, prompt-based classification offers limited insight into decision-making processes, making it difficult to identify and correct systematic errors or biases.

While our research demonstrates the potential for LLMs to achieve comparable performance to human annotators, widespread adoption could lead to displacement of human annotation work. This economic impact should be considered alongside questions of whether automated systems can adequately capture the full spectrum of human interpretive capabilities required for sensitive political content. We acknowledge these ethical considerations and emphasize the importance of responsible development and deployment of automated text classification systems, particularly when applied to politically sensitive content. Future research should incorporate explicit bias mitigation strategies and consider the broader societal implications of automating political discourse analysis.

Acknowledgments

This work is supported by TWON (project number 101095095), a research project funded by the European Union under the Horizon framework (HORIZON-CL2-2022-DEMOCRACY-01-07). The data collection of this work is funded by SOSEC through the Alfred Landecker Foundation.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023.

Falcon-40b: an open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *14th Linguistic Annotation Workshop (LAW 14)*, pages 60–73. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

David J Chalmers. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

DAIR.AI. 2023. Prompt engineering guide.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

António Guterres and UN Secretary-General. 2022. World moving backwards on sustainable development goals, secretary-general tells economic and social council, deploring 'fundamental lack of solidarity'. *Press Release/Secretary-General/Statements and Messages*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292.

Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *CoRR*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Pengfei Liu, 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(9):1–35.

Robert Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, NLP4DH(Digital humanities in languages).

Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

Beren Millidge. 2023. Llms confabulate not hallucinate.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Morris Moscovitch. 1995. Confabulation. *Memory distortions: How minds, brains, and societies reconstruct the past*, page 226–251.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023a. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023b. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, 6(1):4–5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fynn Petersen-Frey, Tim Fischer, Florian Schneider, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023. From qualitative to quantitative research: Semi-automatic annotation scaling in the digital humanities. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 52–62, Ingolstadt, Germany. Association for Computational Lingustics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *The 28th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 1–6.

Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, and Xuan-Jing Huang. 2022. Paradigm shift in natural language processing. *Machine Intelligence Research*, 19(3):169–183.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maximilian Weißenbacher and Udo Kruschwitz. 2023. Steps towards addressing text classification in low-resource languages. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 69–76, Ingolstadt, Germany. Association for Computational Lingustics.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61.

Jules White, 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*. The Hillside Group.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. *Xlnet: Generalized autoregressive pretraining for language understanding*, chapter 1. Curran Associates Inc.