Death of the Dictionary? – The Rise of Zero-Shot Sentiment Classification

Janos Borst, Jannis Klähn and Manuel Burghardt

Computational Humanities, Leipzig University, Germany

Abstract

In our study, we conduct a comparative analysis between dictionary-based sentiment analysis and entailment zero-shot text classification for German sentiment analysis. We evaluate the performance of a selection of dictionaries on eleven data sets, including four domain-specific data sets with a focus on historic German language. Our results demonstrate that, in the majority of cases, zero-shot text classification outperforms general-purpose dictionary-based approaches but falls short of the performance achieved by specifically fine-tuned models. Notably, the zero-shot approach exhibits superior performance, particularly in historic German cases, surpassing both general-purpose dictionaries and even a broadly trained sentiment model. These findings indicate that zero-shot text classification holds significant promise as an alternative, reducing the necessity for domain-specific sentiment dictionaries and narrowing the availability gap of off-the-shelf methods for German sentiment analysis. Additionally, we thoroughly discuss the inherent trade-offs associated with the application of these approaches.

Keywords

sentiment analysis, zero-shot text classification, sentiment dictionary

1. Introduction

Sentiment analysis plays an important role in digital humanities, allowing researchers to uncover attitudes and emotions expressed in text. However, when the text and domain differ from the available datasets, some off-the-shelf methods or models become significantly less useful. Since the data of interest to the humanities often diverge in language and subject from computer science reference datasets, and are rarely fully digitised, let alone annotated, alternative methods that do not require fine-tuning a large language model (LLM) or custom curated dictionary become particularly interesting. We target the domain of historical German language, specifically historical stock market reports and literature, for which there seems to be a lack of readily available domain-specific packages and models.

While there is the established approach of using sentiment dictionaries and the modern approach of fine-tuning LLMs, both lead to significant workloads in aggregating and curating domain-specific data or annotations when deviating from off-the-shelf methods. Recent approaches – namely zero-shot text classification – promise to achieve similar results without manual dataset creation. While fine-tuning neural networks remains the gold standard for

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France
janos.borst@uni-leipzig.de (J. Borst); jannis.klaehn@uni-leipzig.de (J. Klähn); burghardt@informatik.uni-leipzig.de (M. Burghardt)
0000-0002-9166-4069 (J. Borst); 0009-0005-2156-3034 (J. Klähn); 0000-0003-1354-9089 (M. Burghardt)

optimal performance, we explore whether zero-shot sentiment classification can serve as a substitute for the dictionary-based baseline, discussing its advantages and drawbacks.

Current sentiment analysis methods fall into two main categories: dictionary-based and machine learning-based approaches. Dictionaries are the more traditional way to tackle sentiment analysis and is still an actively used approach. In short, the procedure is to use expert knowledge to craft domain- and task-specific lists of negative and positive words with respective sentiment evaluation to build a word-sentiment mapping. These words' occurrences in texts are then aggregated, and their sentiment valuations' ratio or sum determines the text's sentiment index.

While this approach offers advantages like computational efficiency and explainability, it often requires the creation of domain-specific dictionaries or results in performance drops when using general-purpose ones [13]. Many decisions must be made regarding preprocessing, including casing, stemming, and POS-filtering, all of which can impact performance. Additionally, linguistic challenges such as handling negation and metaphors, which are not easily captured in word lists, need consideration. Text quality is also crucial, as the approach requires matching word strings regardless of orthographic or grammatical errors.

The emergence of LLMs such as BERT [16] and GPT [12] has led an the increased popularity of machine learning methods, as they solve many of these problems. The tokenisation process makes them robust against orthographic mistakes, the contextualisation makes it possible to spot negations and contextual semantics. In turn, however, they come with the downside of fine-tuning, requiring substantial manual effort in annotating task-specific data and computational cost to adapt and inference with often over a million parameters.

Consequently, there is a growing interest in exploring more efficient approaches like zero-shot learning [66, 52, 21]. Zero-shot learning offers the potential to automate sentiment analysis tasks by eliminating the need for manual data labeling. Zero-shot learning has already demonstrated promising results in general text classification tasks [67, 21] and application to sentiment analysis [52, 43, 57, 22, 50]. This approach comes with the advantages of robustness against orthographic mistakes and not having to label data, either as training data or word lists, and the capability to detect contextualised semantics. However, it has a larger computational inference time as it still is mostly based on neural networks. This is why we think zero-shot models could be a compromise between the performance of neural language models and close the gap for availability of off-the-shelf methods for sentiment analysis, while keeping the advantages over dictionary-based approaches.

In this paper we try to analyse the performance of these three approaches – a variety of dictionaries, zero-shot-learning and a fine-tuned transformer model – for German sentiment analysis and hope to be able to demonstrate the usefulness of the zero-shot sentiment classification method for application in German language. To get a more valid result we not only test these models on our target domain (historical German) but also on many contemporary German sentiment datasets, such as reviews and tweets.

Our contributions are:

- A comparison of dictionary-based sentiment analysis and zero-shot sentiment classification with regard to performance and inference time.[1]

[1]Code to reproduce results available at https://github.com/JaBorst/deathofthedictionary

- A discussion of advantages and drawbacks of these approaches and their usefulness for practical purposes, with focus on digital humanities datasets.

2. Related Work

Application of dictionary-based sentiment analysis is still an activate field of research [29, 36, 28, 40, 39, 47, 35] with the advent of transformer-based classification, a more sophisticated approach has emerged [27]. The use of computer-assisted text analysis has also become sufficiently established in the humanities and social sciences that performance comparisons of different methods with their own content focus have gained pertinence [9, 1, 62, 2]. A major criticism is that off-the-shelf dictionaries, i.e. existing vocabularies for emotion or trend analysis, are highly domain-dependent in their classification performance [1] and do not provide satisfactory results without revalidation [13]. Furthermore, dictionaries are language-bound and cannot be translated without verification due to the ambiguity of the words they contain.

The prevalence of English dictionaries is a common problem in the field, leading to resource imbalances. In a comparison of different polarity resources in German, [25] found that both quantity and quality differed considerably. Additionally, these manually created sources have proven to be error prone [55]. Moreover, the creation of these annotations is often influenced by domain-specific factors, limiting their generalisability [13, p. 19]. For many use cases, domain-specific dictionaries are required, and while extremely labor-intensive and time-consuming to create, they are still applied in individual cases [28, 40, 39]. However, as [19] show in their comparison of different German dictionaries and datasets, domain-specific dictionaries do not perform well for other applications [40, 19].

Hybrid methods that combine machine learning with semi-automatic word list creation or dictionary expansion have been proposed as promising approaches. These methods are cumbersome due to the cumulative validation steps required [56, 38, 17]. Dictionaries offer the advantage of low-threshold and resource-efficient applicability without requiring training data [47]. Nevertheless, compared to supervised learning methods, both off-the-shelf and specially created dictionaries, including self-implemented and commercial options, consistently show significantly worse performance [5, 9, 17, 1, 62].

In supervised learning, neural networks have emerged as the state-of-the-art for sentiment text classification over the last years. Especially fine-tuning transformer-based LLMs, such as BERT [16], is nowadays the de facto standard in solving text classification tasks [**yangXLNetgeneralizedAutoregressive2019a**, 31]. The main drawback with applying LLMs to new domain-specific tasks is the need for annotated data and the necessary hardware to compute, which can be substantial [49]. Achieving domain-adaptation of LLM-based text classification models through fine-tuning often comes with the computational cost of having to update millions of parameters for every data point, which can be rather difficult and even infeasible at times. In recent years, there has been a significant focus on developing methods that reduce the reliance on large training data sets, leading to the emergence of few-shot models [12, 11, 4, 60] and even zero-shot models [66, 67, 48]. These models enable text classification tasks to be performed without the need for task-specific fine-tuning or manual data labeling. The application of zero-shot text classification models not only eliminates the necessity for manual data annotation but also mitigates the computational costs associated with fine-tuning. Therefore, we systematically investigate the performance of zero-shot against dictionaries for the task of sentiment analysis on German texts for both general and domain-specific use cases.

3. Experiments

In this section we briefly describe the experimental setting. We explain the application of the dictionaries and zero-shot methods and list the datasets we used to compare them.

3.1. Dictionaries

In order to obtain fair comparisons for German dictionaries, we decided upon three generally applicable German off-the-shelf-dictionaries (SentiWS, BAWL-R, GermanPolarityClues) with a wide reputation [41, 58, 59]. In addition, a finance-specific dictionary BPW was tested for the special dataset BBZ [3], as well as a literature-specific dictionary SentiLitKrit (SLK)[18]. Both SentiWS and BAWL-R offer valence-based sentiment classification, meaning that each word in the dictionary is weighted by a numerical value, whereas the other dictionaries only allow for polarity-based sentiment assignment. For the annotation of the datasets with the presented off-the-shelf-dictionaries, we follow the approach of [1], [5], as well as [62], who all use the R library quanteda and a similar pre-processing. In our case, the quanteda extension quanteda sentiment was used and only punctations and numbers were removed.[2]

3.2. Zero-Shot Text Classification

As a zero-shot model, we use textual entailment classification – also called natural language inference (NLI) –, following the task description proposed in [67]. In this approach a sentence pair, called premise and hypothesis, is classified as 'entailment', 'contradiction' or 'neutral', based on how well the hypothesis logically entails the premise. For zero-shot classification we form hypotheses using the target labels. These hypotheses are created using the template: *"The sentiment is [blank]"*[3]. The blank is then filled with the sentiment categories 'negative', 'neutral' and 'positive'. The 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. This methodology is applied to achieve zero-shot sentiment classification, as illustrated in Figure 1.

Although there is some criticism about the performance of these models relying on spurious correlation of superficial text elements [33], still this model – and variants of it – are performing very well, especially in sentiment classification [69, 52]. We choose this the entailment

[2]https://github.com/quanteda/quanteda.sentiment. Besides the simpler quanteda variant, there are also more complex dictionary approaches, such as VADER [23]. However, as VADER was developed for the English language and the integrated translation option would not be cost-free for the datasets tested here, it was decided not to use VADER.

[3]Translated from German: *"Die Stimmung ist [blank]."*

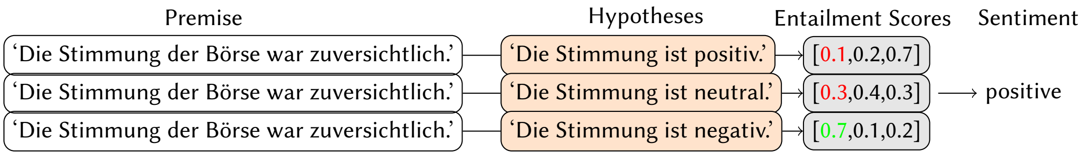

Figure 1: Step-by-step example of classifying a single example in an entailment-based zero-shot sentiment classification. Premise and hypothesis are concatenated into one string, which is classified into the entailment classes, neutral and contradiction. The entailment scores are then compared within the set of hypotheses. The hypothesis with the highest entailment score is then chosen to be the classification result.

approach also because of its flexibility and accessibility. As entailment model we use a pre-trained NLI-model[4] from huggingface[65], which was trained on machine-translated versions of multiple NLI-datasets (MNLI [63]), ANLI [37], SNLI [10]) and tested on the German part of the XNLI [15] dataset.

Since we aim at comparing these models with zero knowledge about domain specific assumptions or vocabulary, we use the same for all datasets. For application purposes the hypothesis template can have substantial impact on the quality of classification, and is part of the optimization process similar to prompt engineering [30]. The problem with optimizing the hypothesis template is the need for some annotated data for evaluation, which is another reason, why we opt for a simple, fixed and generic template.

3.3. Finetuned or State-of-the-Art

As for the finetuned and state-of-the-art performance on each dataset we include two comparisons: Firstly, we include the latest developments in the field as reported in other papers, showcasing the current state-of-the-art (SOTA). Secondly, we employ a German sentiment model developed by [20]. This off-the-shelf model is trained on various German sentiment datasets and serves as another benchmark for comparison. It is worth exploring whether this broadly trained model, without domain-specific adaptation, generalises well on out-of-domain datasets, e.g. German historic language.

3.4. Data

We chose the datasets by availability and by recent mentions of SOTA results in recent research publications. Nevertheless, the availability of non-English datasets remains a severely limiting factor and also poses a significant problem for the subsequent training of specialised language models. In addition to seven contemporary datasets based on social media posts or reviews, we also selected four domain specific datasets based on historical German, which we will outline below:

[4] svalabs/gbert-large-zero-shot-nli

BBZ [61]: A set of 772 sentences sampled from the Berliner Börsenzeitung (BBZ) between 1872 and 1930. The dataset contains sentences-level annotations of negative, neutral and positive labels.

German Novel Dataset (GND) [68]: A crowd-sourced collection of 270 ternary labelled sentences (positive,neutral,negative) from the German Novel Corpus (GNC).

Lessing [46]: A set of 200 sampled speeches from Gottlob Lessing's plays, manually annotated by five experts with binary labels (positive, negative).

SentiLitKrit [18]: A sample corpus for the SentiLitKrit (SLK) dictionary consisting of manually annotated literature reviews for the period 1870-1889 with 1,010 binary annotated sentences.

GermEval 2017 [64]: This collection of tweets and news about the Deutsche Bahn between 2015 and 2016. We use the predefined synchronous test set with 2,566 examples labelled with positive, neutral or negative.

PotTs [53]: A collection of tweets from 2013 on elections and political events with 7,504 items. The labels are positive, neutral or negative.

SB10k [14]: Originally a set of over 9,000 tweets collected in 2013 divided into the categories positive, negative and neutral. Since the original dataset does only publish the Twitter links, we resort to a pre-assembled version by [20] with 7,476 entries and the labels positive, neutral or negative.

Amazon Reviews DE [26]: A multilingual corpus of amazon product reviews based on star-ratings between 2015 and 2019. We use the German part of this corpus, which contains 5,000 test set elements.

Filmstarts and Holidaycheck [20]: These datasets are sets of reviews for either films or hotels crawled form the respective website. We use the dataset as described by [20] to ensure comparability and also exclude ratings with 3 stars, which would correspond to the label neutral. The resulting sets include 55,260 items for Filmstarts and around 3.3 million items for Holidaycheck.

SCARE [42]: This dataset contains around 735,000 reviews for various apps from the Google Playstore. It contains positive, neutral or negative labels.

Except for the Amazon reviews, all datasets are unbalanced. Unless stated otherwise, no pre-processing was conducted and if no dedicated test dataset was available, the entire dataset was annotated. You can find detailed table of dataset sizes and composition in the Appendix A.

4. Results & Discussion

Performance In Table 1, the micro F1 evaluation scores for all datasets and approaches are presented. It is important to note that when comparing against [20], a problem arises, as we were unable to reproduce the exact test sets used in their reported values. Therefore, there is a possibility that our evaluation of their model may differ from the SOTA value extracted from their work. Furthermore, to reflect the criticism of the high technical barrier faced by social scientists, an out-of-the-box approach of implementing the Guhr et al. model using the Huggingface pipeline was applied.

The experimental findings indicate a consistent pattern in the performance of zero-shot text classification, which falls between the application of available dictionaries and the SOTA approach in micro and macro F1 (Table 1 and Table 2). This pattern holds true not only for contemporary data such as Amazon reviews or tweets but also generalises to the historical examples like BBZ, GND, SentiLitKrit and Lessing. Close inspection into label-wise metrics as seen in Table 3 reveals that this happens despite of zero-shot struggling with the neutral class. This also explains its high performance in binary polarity cases. The performance on positive and negative polarity is high, with the exception of SB10k and GermEval, this will be discussed shortly below.

Table 1
Micro F1 for all approaches on all datasets. Best scores per dataset are marked in bold, excluding the state of the art.

Dataset	SLK	BPW	BAWL-R	SentiWS	GPC	Zero-shot	Guhr et al.	SOTA
BBZ[61]	0.371	0.435	0.371	0.519	0.511	**0.676**	0.272	0.884[8]
GND[68]	**0.496**	0.474	0.348	0.437	0.455	0.466	0.448	0.430[68]
Lessing[46]	0.387	0.557	0.424	0.582	0.608	**0.746**	0.506	0.627[44]
SentiLitKrit[18]	0.662	0.515	0.652	0.665	0.621	**0.787**	0.111	0.76[18]
GermEval2017[64]	0.485	0.563	0.242	0.380	0.366	0.332	**0.583**	0.851[24]
PotTs [53]	0.388	0.406	0.431	0.461	0.437	**0.516**	0.389	0.650[20]
SB10k [14]	0.539	0.487	0.369	0.365	0.435	0.335	**0.614**	0.773[6]
Amazon Reviews[26]	0.425	0.464	0.416	0.581	0.582	**0.697**	0.669	0.734[34]
Filmstarts [20]	0.674	0.596	0.703	0.743	0.719	0.822	**0.831**	0.921[20]
Holiday Check [20]	0.649	0.696	0.844	0.853	0.824	0.929	**0.935**	0.977[20]
SCARE [42]	0.315	0.368	0.494	0.722	0.726	**0.879**	0.797	0.943[20]

Table 2
Macro F1 for all approaches on all datasets. Best scores per dataset are marked in bold, excluding the state of the art. Macro F1 gets reported less often, which is why in most cases the SOTA column is empty.

Dataset	SLK	BPW	BAWL-R	SentiWS	GPC	Zero-shot	Guhr et al.	SOTA
BBZ[61]	0.368	0.437	0.348	0.465	0.475	**0.642**	0.167	0.807[8]
GND[68]	0.483	0.446	0.342	0.429	**0.450**	0.438	0.322	-
Lessing[46]	0.385	0.481	0.415	0.557	0.585	**0.712**	0.388	-
SentiLitKrit[18]	0.585	0.506	0.503	0.613	0.573	**0.759**	0.137	0.76[18]
GermEval2017[64]	0.328	0.437	0.235	0.349	0.334	0.307	**0.442**	-
PotTs [53]	0.322	0.365	0.393	**0.454**	0.435	0.452	0.355	-
SB10k [14]	0.356	0.355	0.320	0.356	0.390	0.351	**0.513**	0.748[7]
Amazon Reviews[26]	0.414	0.452	0.340	0.467	0.510	**0.616**	0.548	-
Filmstarts [20]	0.597	0.579	0.476	0.692	0.671	0.774	**0.816**	-
Holiday Check [20]	0.509	0.581	0.505	0.654	0.695	0.843	**0.870**	-
SCARE [42]	0.301	0.389	0.416	0.666	0.684	**0.856**	0.779	-

While the off-the-shelf model by [20] achieves a slightly better result than zero-shot classification on contemporary data, it fails to generalise effectively to the historical and literature

Table 3
Precision, recall and micro F1-Values for each label for the zero-shot approach on all datasets.

Dataset	Negative	Neutral	Positive
BBZ Gold	0.698 | 0.642 | 0.669	0.469 | 0.545 | 0.504	0.821 | 0.792 | 0.807
Lessing [46]	0.853 | 0.755 | 0.801	-	0.558 | 0.704 | 0.623
SentiLitKrit [18]	0.603 | 0.770 | 0.676	-	0.894 | 0.793 | 0.841
GND [68]	0.475 | 0.775 | 0.589	0.700 | 0.112 | 0.194	0.409 | 0.754 | 0.530
GermEval2017 [64]	0.503 | 0.767 | 0.608	0.769 | 0.097 | 0.173	0.076 | 0.847 | 0.140
PotTs [53]	0.425 | 0.806 | 0.557	0.415 | 0.109 | 0.172	0.599 | 0.673 | 0.634
SB10k [14]	0.318 | 0.746 | 0.446	0.687 | 0.076 | 0.138	0.327 | 0.822 | 0.468
Amazon Reviews [26]	0.744 | 0.769 | 0.757	0.358 | 0.244 | 0.290	0.756 | 0.852 | 0.801
Filmstarts [20]	0.649 | 0.797 | 0.715	-	0.913 | 0.831 | 0.870
Holiday Check [20]	0.663 | 0.805 | 0.727	-	0.973 | 0.946 | 0.959
SCARE [42]	0.741 | 0.845 | 0.789	-	0.940 | 0.891 | 0.915

Table 4
Time measurements in items/s, averaged over all datasets, to compare run times between these approaches. The dictionaries are measured on a standard CPU while the neural networks are run on a RTX 2080 Ti.

Method	BPW	SLK	BAWL-R	SentiWS	GPC	Zero-shot	Guhr et al.
Time (items/s)	50.756	83.639	57.394	18.285	52.590	49	201

domain. Given that the model was trained on contemporary or similar domain and language style datasets, this is not surprising, but also illustrates that even modern language models without fine-tuning in the envisaged target domain only achieve mediocre results.

Generally, our tests show a strong inconsistency in the results of the dictionaries, which is independent of the intended use. In the two cases, GermEval 2017 and SB10k, where zero-shot performs worse than dictionaries, we see a pattern of texts of very low quality. These datasets contain annotations of varying quality and appear to be somewhat inconsistent. This is why being trained on this type of data grants substantial advantage. However, the dictionary approach, especially using BPW, although designed for financial contexts, seems to be working well in these case. Moreover, a larger vocabulary or a combination of dictionaries, such as the 2012 version of GPC, which also contains the SentiWS vocabulary, does not necessarily lead to better results. In the case of the SLK dataset with the purpose-built dictionary, the enormous effort required to create the dictionary is not reflected in a significantly better performance, as can be seen in Table 1.

For the Lessing dataset, an additional argument can be made regarding the inherent incoherence of sentiment annotations. The ambiguity in sentiment often leads to low inter-annotator agreement during the annotation process [45]. In this context, the zero-shot algorithm demonstrates its effectiveness by aligning with the majority decision in determining sentiment.

Nevertheless based on these performance observations, we argue that the results provide evidence supporting the viability of zero-shot text classification as a potential alternative, if not a replacement, for general-purpose polarity dictionaries. Particularly in use cases where

Table 5
An assessment of various aspects and trade-offs when comparing off-the-shelve dictionary, zero-shot and trained models.

	Dictionary	Zero-shot	Finetuning
Preprocessing	-	+	+
Robustess	-	+	+
Domain Adaptation	-	+	-
Inference Time	+	-	-
Performance	-	+	+
Explainability	+	o	o
Hardware	+	-/o	-

no annotated training data or domain-specific dictionaries are available, but where the linguistic complexity or subject matter is different from that of the existing general-purpose models/dictionaries, the zero-shot approach presented here delivers significantly better results and a higher consistency of performance, provided that the quality of the source text is not too low.

Trade-offs As is often the case, there are several trade-offs to consider. In Table 5 we mark these trade-offs for the methods with -, o and +, denoting disadvantage, neutral or advantage. LLM tokenisation eases preprocessing and enhances robustness against orthographic errors and contextual semantics, issues dictionary-based methods struggle with. Adapting dictionaries or LLMs to specific domains can be costly. Zero-shot models hold a clear advantage due to their flexibility without adaptation. In cases where dictionaries are not adapted to the specific domain, the entailment zero-shot approach would deliver better performance in most cases. Fine-tuning of the language models will deliver the best performance in any case, if trained for the specific task. The dictionary approach takes the clear win in inference time and explainability. During inference time, entailment zero-shot and fine-tuning are both slower than dictionaries. The factor of around 4x (3x for shorter texts) between zero-shot and Guhr et al. stems from the fact that entailment formulation introduces a forward pass per label, which in our case is two or three, and the base model for zero-shot has three times the parameters (109M vs 330M).

Dictionaries offer clear explanations for algorithmic decision-making, directly tracking each word's contribution to sentiment scores. However, performance may not align with this theoretical comprehensibility, as indicated in the evaluation of the financial BPW dictionary. In contrast, neural classifiers are often regarded as black boxes, but there are ongoing efforts to explain token influences on classification results [**NIPS20178a20a862** , 54, 51], albeit through mathematical approximations. Since this is a more indirect measure, we assess this as neutral (o) for now.

Another point to consider is that the inference and also training time, if necessary, depend strongly on the hardware used. While the dictionary approaches are very efficient and do not need special hardware, neural network based classifiers often gain speed significantly from using GPUs, with the limiting factor often being the VRAM. Luckily, during inference the requirements are a bit lower than during training and especially the model we used is able to run easily on consumer-grade GPUs [70].

5. Conclusion

In our study, we conducted a comparative analysis of three approaches to German sentiment analysis: dictionary-based, zero-shot, and fine-tuning. Although there are certain trade-offs, the viability of zero-shot text classification for sentiment analysis as a possible alternative to dictionary-based methods can be reasonably argued, particularly in cases where a fine-tuned model cannot be applied or trained sufficiently, either because of a lack of training data or due to more specific domains that deviate from the standard approaches trained on tweets or reviews. Especially in binary cases there seem to be a clear advantage of applying zero-shot models to alleviate data labeling labour with still substantial performance. We also emphasise that this paper was not concerned with fine-tuning or further engineering the prompt: In future work the zero-shot's weakness for neutral labels could be a matter of designing a better hypothesis template.

We argue that zero-shot text classification for polarity sentiment could also contribute to bridging the gap in model availability for languages other than English. In our research, we specifically focused on an entailment-based zero-shot approach. However, with the introduction of advanced language models like GPT-4 or LLama, the performance of zero-shot text classification is expected to further widen the gap between dictionary approaches and zero-shot text classification and even bring zero-shot results closer to SOTA values.

Acknowledgements

This research is funded by the German Research Foundation (DFG) under project number BU 3502/1-1.

References

[1] W. van Atteveldt, M. A. C. G. van der Velden, and M. Boukes. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms". In: *Communication Methods and Measures* 15.2 (2021), pp. 121–140. DOI: 10.1080/19312458.2020.1869198.

[2] C. Baden, C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda". In: *Communication Methods and Measures* 16.1 (2022), pp. 1–18. DOI: 10.1080/19312458.2021.2015574.

[3] C. Bannier, T. Pauls, and A. Walter. "Content analysis of business communication: introducing a German dictionary". In: *Journal of Business Economics* 89.1 (2019), pp. 79–123. DOI: 10.1007/s11573-018-0914-8.

[4] Y. Bao, M. Wu, S. Chang, and R. Barzilay. "Few-shot Text Classification with Distributional Signatures". In: *International Conference on Learning Representations*. 2020,

[5] P. Barberá, A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler. "Automated Text Classification of News Articles: A Practical Guide". In: *Political Analysis* 29.1 (2021), pp. 19–42. DOI: 10.1017/pan.2020.8.

[6] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 258–266.

[7] V. Barriere and A. Balahur. "Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, and C. Zong. Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 266–271. DOI: 10.18653/v1/2020.coling-main.23.

[8] J. Borst, L. Wehrheim, and M. Burghardt. ""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*. Ed. by A. Baillot, T. Tasovac, W. Scholger, and G. Vogeler. 2023, pp. 365–367.

[9] M. Boukes, B. van de Velde, T. Araujo, and R. Vliegenthart. "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools". In: *Communication Methods and Measures* 14.2 (2020), pp. 83–104. DOI: 10.1080/19312458.2019.1671966.

[10] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075.

[11] J. Bragg, A. Cohan, K. Lo, and I. Beltagy. "FLEX: Unifying Evaluation for Few-Shot NLP". In: *NeurIPS 2021*. 2021,

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[13] C.-H. Chan, J. Bajjalieh, L. Auvil, H. Wessler, S. Althaus, K. Welbers, W. van Atteveldt, and M. Jungblut. "Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: a large-scale p-hacking experiment". In: *Computational Communication Research* 3.1 (2021), pp. 1–27. URL: https://computationalcommunication.org/ccr/article/view/40.

[14] M. Cieliebak, J. M. Deriu, D. Egger, and F. Uzdilli. "A Twitter Corpus and Benchmark Resources for German Sentiment Analysis". In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Ed. by L.-W. Ku and C.-T. Li. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 45–51. DOI: 10.18653/v1/W17-1106.

[15] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. "XNLI: Evaluating cross-lingual sentence representations". In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2018,

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* (2019). DOI: {arXiv}:1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805.

[17] T. Dobbrick, J. Jakob, C.-H. Chan, and H. Wessler. "Enhancing Theory-Informed Dictionary Approaches with "Glass-box" Machine Learning: The Case of Integrative Complexity in Social Media Comments". In: *Communication Methods and Measures* 16.4 (2022), pp. 303–320. DOI: 10.1080/19312458.2021.1999913.

[18] K. Du and K. Mellmann. *Sentimentanalyse als Instrument literaturgeschichtlicher Rezeptionsforschung*. Working Paper. Göttingen, 2019.

[19] J. Fehle, T. Schmidt, and C. Wolff. *Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques.* 2021. DOI: 10.5283/epub.50833.

[20] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme. "Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, 2020, pp. 1627–1632.

[21] K. Halder, A. Akbik, J. Krapac, and R. Vollgraf. "Task-Aware Representation of Sentences for Generic Text Classification". In: *Proceedings of the 28th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, 2020, pp. 3202–3213. DOI: 10.18653/v1/2020.coling-main.285.

[22] M. Hu, S. Zhao, H. Guo, C. Xue, H. Gao, T. Gao, R. Cheng, and Z. Su. "Multi-Label Few-Shot Learning for Aspect Category Detection". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, 2021, pp. 6330–6340. DOI: 10.18653/v1/2021.acl-long.495.

[23] C. Hutto and E. Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (2014), pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[24] A. Idrissi-Yaghir, H. Schäfer, N. Bauer, and C. M. Friedrich. "Domain Adaptation of Transformer-Based Models Using Unlabeled Data for Relevance and Polarity Classification of German Customer Feedback". In: *SN Computer Science* 4.2 (2023). DOI: 10.1007/s42979-022-01563-6.

[25] B. M. J. Kern, A. Baumann, T. E. Kolb, K. Sekanina, K. Hofmann, T. Wissik, and J. Neidhardt. *A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis.* 2021. DOI: 10.4230/oasics.ldk.2021.37.

[26] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith. “The Multilingual Amazon Reviews Corpus”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 4563–4568. DOI: 10.18653/v1/2020.emnlp-main.369.

[27] E. Kim and R. Klinger. *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies*. 2019. DOI: 10.17175/2019{\textunderscore}008.

[28] T. Kolb, Sekanina Katharina, B. M. J. Kern, J. Neidhardt, T. Wissik, and A. Baumann. “The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*. Ed. by N. e. a. Calzolari. 2022, pp. 4708–4716.

[29] S. Lee, S. Ma, J. Meng, J. Zhuang, and T.-Q. Peng. “Detecting Sentiment toward Emerging Infectious Diseases on Social Media: A Validity Evaluation of Dictionary-Based Sentiment Analysis”. In: *International Journal of Environmental Research and Public Health* 19.11 (2022). DOI: 10.3390/ijerph19116759.

[30] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. “Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. In: *ACM Comput. Surv.* 55.9 (2023). DOI: 10.1145/3560815.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[32] S. M. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017,

[33] T. Ma, J.-G. Yao, C.-Y. Lin, and T. Zhao. “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)*. Association for Computational Linguistics, 2021, pp. 786–796. DOI: 10.18653/v1/2021.acl-short.99.

[34] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis. “Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data”. In: *Neural Computing and Applications* (2023). DOI: 10.1007/s00521-023-08629-3.

[35] A. Mengelkamp, K. Koch, and M. Schumann. “Creating Sentiment Dictionaries: Process Model and Quantitative Study for Credit Risk”. In: *Proceedings of the 9th European Conference on Social Media*. 1. 2022, pp. 121–129. DOI: 10.25968/opus-2449.

[36] K. Müller. “German forecasters’ narratives: How informative are German business cycle forecast reports?” In: *Empirical Economics* 62.5 (2022), pp. 2373–2415. DOI: 10.1007/s00181-021-02100-9.

[37] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. “Adversarial NLI: A New Benchmark for Natural Language Understanding”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 4885–4901. DOI: 10.18653/v1/2020.acl-main.441.

[38] M. Palmer, J. Roeder, and J. Muntermann. “Induction of a sentiment dictionary for financial analyst communication: a data-driven approach balancing machine learning and human intuition”. In: *Journal of Business Analytics* 5.1 (2022), pp. 8–28. DOI: 10.1080/2573234x.2021.1955022.

[39] M. Pöferlein. “Sentiment Analysis of German Texts in Finance: Improving and Testing the BPW Dictionary”. In: *Journal of Banking and Financial Economics* 2021.2(16) (2021), pp. 5–24. DOI: 10.7172/2353-6845.jbfe.2021.2.1.

[40] C. Puschmann, H. Karakurt, C. Amlinger, N. Gess, and O. Nachtwey. “RPC-Lex: A dictionary to measure German right-wing populist conspiracy discourse online”. In: *Convergence (London, England)* 28.4 (2022), pp. 1144–1171. DOI: 10.1177/13548565221109440.

[41] R. Remus, U. Quasthoff, and G. Heyer. “SentiWS - A Publicly Available German-language Resource for Sentiment Analysis”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010, p. 2.

[42] M. Sänger, U. Leser, S. Kemmerer, P. Adolphs, and R. Klinger. “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)*. Ed. by N. C. (Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Paris, France: European Language Resources Association (ELRA), 2016,

[43] A. Sarkar, S. Reddy, and R. S. Iyengar. “Zero-Shot Multilingual Sentiment Analysis Using Hierarchical Attentive Network and BERT”. In: *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*. Nlpir 2019. New York, NY, USA: Association for Computing Machinery, 2019, pp. 49–56.

[44] T. Schmidt and M. Burghardt. “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*. Ed. by B. Alex. Stroudsburg, PA: Association for Computational Linguistics, 2018, pp. 139–149.

[45] T. Schmidt, M. Burghardt, and K. Dennerlein. “„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*. 2018,

[46] T. Schmidt, M. Burghardt, and K. Dennerlein. “Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior”. In: *Proceedings of the Workshop on Annotation in Digital Humanities 2018 (annDH 2018)*. Ed. by S. Kübler and H. Zinsmeister. Sofia, Bulgaria, 2018, pp. 47–52.

[47] T. Schmidt, J. Dangel, and C. Wolff. *SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities*. Universität Regensburg, 2021. DOI: 10.5283/epub.44943.

[48] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. “Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2019, pp. 8239–8247. DOI: 10.1109/cvpr.2019.00844.

[49] R. Schwartz, J. Dodge, N. Smith, and O. Etzioni. “Green AI”. In: *Communications of the ACM* 63 (2019), pp. 54–63.

[50] R. Seoh, I. Birle, M. Tak, H.-S. Chang, B. Pinette, and A. Hough. “Open Aspect Target Sentiment Classification with Natural Language Prompts”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 6311–6322. DOI: 10.18653/v1/2021.emnlp-main.509.

[51] A. Shrikumar, P. Greenside, and A. Kundaje. “Learning important features through propagating activation differences”. In: *Proceedings of the 34th international conference on machine learning - volume 70*. Icml’17. JMLR.org, 2017, pp. 3145–3153.

[52] L. Shu, H. Xu, B. Liu, and J. Chen. *Zero-Shot Aspect-Based Sentiment Analysis*. 2022. arXiv: 2202.01924 [cs.CL].

[53] U. Sidarenka. “PotTS: The Potsdam Twitter Sentiment Corpus”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016, pp. 1133–1141.

[54] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *Workshop at International Conference on Learning Representations*. 2014,

[55] H. Song, P. Tolochko, J.-M. Eberl, O. Eisele, E. Greussing, T. Heidenreich, F. Lind, S. Galyga, and H. G. Boomgaarden. “In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis”. In: *Political Communication* 37.4 (2020), pp. 550–572. DOI: 10.1080/10584609.2020.1723752.

[56] A. Stoll, L. Wilms, and M. Ziegele. “Developing an Incivility Dictionary for German Online Discussions – a Semi-Automated Approach Combining Human and Artificial Knowledge”. In: *Communication Methods and Measures* (2023), pp. 1–19. DOI: 10.1080/19312458.2023.2166028.

[57] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, and R. Damaševičius. “Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning”. In: *Applied Sciences* 12.17 (2022), p. 8662.

[58] M. L. H. Võ, M. Conrad, L. Kuchinke, K. Urton, M. J. Hofmann, and A. M. Jacobs. “The Berlin Affective Word List Reloaded (BAWL-R)”. In: *Behavior Research Methods* 41.2 (2009), pp. 534–538. DOI: 10.3758/brm.41.2.534.

[59] U. Waltinger. “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, 2010, p. 00.

[60] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. “Generalizing from a Few Examples: A Survey on Few-Shot Learning”. In: *ACM Comput. Surv.* 53.3 (2020). DOI: 10.1145/3386252.

[61] L. Wehrheim, J. Borst, M. Burghardt, and A. Niekler. “„Auch heute war die Stimmung im Allgemeinen fest.“ Zero-Shot Klassifikation zur Bestimmung des Media Sentiment an der Berliner Börse zwischen 1872 und 1930”. In: *Konferenzabstracts DHd2023: Open Humanities, Open Culture*. 2023, pp. 90–94.

[62] T. Widmann and M. Wich. “Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text”. In: *Political Analysis* (2022), pp. 1–16. DOI: 10.1017/pan.2022.15.

[63] A. Williams, N. Nangia, and S. Bowman. “A broad-coverage challenge corpus for sentence understanding through inference”. In: *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122.

[64] M. Wojatzki, E. Ruppert, S. Holschneider, T. Zesch, and C. Biemann. “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*. Berlin, Germany, 2017, pp. 1–12.

[65] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

[66] Y. Xian, B. Schiele, and Z. Akata. “Zero-Shot Learning - The Good, the Bad and the Ugly”. In: *IEEE Computer Vision and Pattern Recognition (CVPR)*. 2017,

[67] W. Yin, J. Hay, and D. Roth. “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*. Association for Computational Linguistics, 2019, pp. 3912–3921. DOI: 10.18653/v1/D19-1404.

[68] A. Zehe, M. Becker, F. Jannidis, and A. Hotho. “Towards Sentiment Analysis on German Literature”. In: *KI 2017: Advances in Artificial Intelligence*. Ed. by G. Kern-Isberner, J. Fürnkranz, and M. Thimm. Vol. 10505. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 387–394. DOI: 10.1007/978-3-319-67190-1_36.

[69] R. H. Zhang, A. X. Fan, and R. Zhang. *ConEntail: An Entailment-based Framework for Universal Zero and Few Shot Classification with Supervised Contrastive Pretraining.* Dubrovnik, Croatia, 2023. DOI: 10.18653/v1/2023.eacl-main.142. URL: https://aclanthology.org/2023.eacl-main.142.

[70] F. Ziegner, J. Borst, A. Niekler, and M. Potthast. *Using Language Models on Low-end Hardware.* 2023. arXiv: 2305.02350 [cs.CL].

A. Data Set Sizes

Appendix A shows an exact breakdown of how many positive, negative and neutral data sets are in each data set.

Table 6
Items per label of all data sets.

Dataset	Negative	Neutral	Positive	Total
BBZ Gold	260	198	314	772
Lessing[46]	139	-	61	200
SentiLitKrit[18]	292	-	718	1010
GND[68]	89	124	57	270
GermEval 2017 sync [64]	780	1681	105	2566
PotTs [53]	1569	2487	3448	7504
SB10k [14]	1130	4629	1717	7476
Amazon Reviews german [26]	2000	1000	2000	5000
Filmstarts [20]	15608	-	40012	55620
Holiday Check [20]	379683	-	2871076	325079
SCARE [42]	196953	-	537629	734592