DOI: 10.15439/2023F3725
ISSN 2300-5963 ACSIS, Vol. 37

An Evaluation of a Zero-Shot Approach to Aspect-Based Sentiment Classification in Historic German Stock Market Reports

Janos Borst*, Lino Wehrheim†, Andreas Niekler*, Manuel Burghardt*
*Leipzig University, Email: [name].[surname]@uni-leipzig.de
†University of Regensburg, Lino.Wehrheim@ur.de

***Abstract*—One critical aspect that remains in the application of state-of-the-art neural networks to text analysis in applied research is the continued requirement for manual data annotation. In computer science research, there is a strong focus on maximizing the data efficiency of fine-tuning language models. This has led to the development of zero-shot text classification methods, which promise to work effectively without requiring fine-tuning for the specific task at hand. In this paper, we conduct an in-depth analysis of aspect-based sentiment analysis in historic German stock market reports to evaluate the reliability of this promise. We present a comparison of a zero-shot approach with a meticulously fine-tuned three-step process of training and applying text classification models. This study aims to empirically assess the reliability of zero-shot text classification and provide justification for the potential benefits it offers in terms of reducing the burden of data labeling and training for analysis purposes. The findings of our study demonstrate a strong correlation between the sentiment time series generated through aspect-based sentiment analysis using the zero-shot approach and those derived from the fine-tuned supervised pipeline, validating the viability of the zero-shot approach. While the zero-shot pipeline exhibits a tendency to underestimate negative examples, the overall trend remains discernible. Additionally, a qualitative analysis of the linguistic patterns reveals no explicit error patterns. Nevertheless, we acknowledge and discuss the practical and epistemological obstacles associated with employing zero-shot algorithms in untested domains.**

I. Introduction

SENTIMENT analysis plays a crucial role in the field of digital humanities, enabling researchers to uncover attitudes and emotions expressed in various forms of text. Existing sentiment analysis approaches can be broadly categorized into dictionary-based methods and machine learning-based methods [1].

With the advent of large language models like BERT [2] or GPT [3], machine learning approaches have gained popularity due to their ability to be fine-tuned rather than trained from scratch. However, the process of fine-tuning models remains laborious and time-consuming, demanding significant manual effort. As a result, there is a growing interest in exploring alternative approaches such as zero-shot learning [4, 5, 6] for sentiment analysis, particularly aspect-based sentiment analysis [5]. Zero-shot learning eliminates the need for manual data labeling, offering a promising avenue for automating sentiment analysis tasks. While zero-shot learning has shown promising results for general text classification tasks already [7, 6], and it also has been tested for sentiment analysis tasks specifically [8, 9, 10, 11, 5], its practical application in digital humanities (DH) projects remains relatively scarce.

To encourage more use of zero-shot approaches in DH, we present a first study that systematically evaluates the effectiveness of zero-shot text classification for aspect-based sentiment analysis. Our evaluation design is inspired by an ongoing research project called “More than a Feeling: Media Sentiment as a Mirror of Investors’ Expectations at the Berlin Stock Exchange, 1872-1930”, which is focused on detecting sentiment in historical German stock market reports. This research project serves as an exemplary case within the realm of digital humanities, highlighting the significant challenges associated with historic sources and languages. To provide a comprehensive evaluation, we compare the performance of the zero-shot approach against another machine learning approach that relies on manually annotated training data to fine-tune existing large language models. This approach was carried out in the initial phase of the above research project [12, 13] and now serves as a baseline to assess the quality of the fully automatic zero-shot approach.

By conducting this systematic evaluation, we aim to contribute to the understanding of zero-shot text classification for aspect-based sentiment analysis, thereby paving the way for its wider application in digital humanities research. Furthermore, we seek to address the unique challenges posed by historic sources and languages, enriching the discourse on sentiment analysis in the context of historical texts. The contributions of this paper are as follows:

- An in-depth evaluation of a zero-shot text classification pipeline for aspect-based sentiment analysis on historical German text data
- In-depth comparison of zero-shot text classification and trained text classification on a complex research application in a quantitative and qualitative manner.
- Insights into and discussion of the potential and limitations of this approach.

II. Related Work

Sentiment analysis is widely used in the field of text mining and social media analytics. In recent years, it has also gained increasing popularity in the Digital Humanities, particularly in the field of Computational Literary Studies [14]. Another

Thematic track: International Workshop on AI in Digital Humanities, Computational Social Sciences and Economics Research

field that is particularly fond of sentiment analysis is Finance and Financial Economics. In fact, it has long been known that economies are heavily influenced by moods, feelings and emotions [15]. Sentiment analysis in financial texts has first been approached by dictionary-based methods [16, 17], which are still used today in some cases [18]. Since machine learning approaches emerged, Transformers have been adapted and applied [19, 20]. Accordingly, resources such as FinBERT [21, 22] are also publicly available for application. One limitation here is that FinBERT only works with texts in the English language. Sentiment classification in these existing works is mainly regarded as sentence classification tasks. However, Sinha et al. [23] note that sentiment in these texts are often specifically entity-related, which can complicate analysis considerably. This challenge also applies to our paper, since the corpus often includes statements referring to entities at different granularity with contrary sentiment valuations, which we will explain later on. This is why we regard the sentiment analysis as an aspect-based sentiment text classification task [24].

Text classification and natural language processing (NLP) in general have made significant progress in recent years. In particular the accessibility of pretrained large language models (LLM) like BERT [2] through Huggingface [25] has had considerable impact on applications. While virtually every metric in NLP has jumped up by employing today's de-facto standard of finetuning LLMs [2, 26, 27], this comes with two caveats: Computational efficiency and data efficiency. While pretraining models has significantly reduced the amount of data needed to achieve competitive results, fine-tuning LLMs often comes with the computational cost of having to update billions of parameters, which can be rather difficult and even infeasible at times. In recent years, research has concentrated on methods that decrease the number of data points needed for training, leading to so-called few-shot models [3, 28, 29, 30] and even zero-shot models [4, 7, 31]. Zero-shot text classification models can be applied to text classification tasks without the need for task-specific fine-tuning or manual data labeling. This alleviates not only the the need for manual data annotation, but also the corresponding computational costs. The formulation of zero-shot text classification as an entailment of sentence pairs [7] serves as a very flexible approach that even can be adapted to aspect-based sentiment classification [5]. It has shown promising results in both sentiment and aspect-based sentiment classification [5]. Another way to apply sentiment analysis to a corpus without having to fine-tune is to make use of publicly shared trained sentiment models. There exists a broadly trained off-the-shelf solutions for German sentiment analysis text [32], which marks an inbetween of models that are trained for the task, but not specifically fine-tuned with domain data.

While there has been some work regarding zero-shot entity recognition in historic German newspaper [33], to the best of our knowledge, we are the first to apply zero-shot aspect-based sentiment classification to German texts. We present an in-depth comparison between the zero-shot approach and specifically trained models, fine-tuned on hand-coded data, for the application on historic German texts.

III. Approach

A. Introduction to the Corpus

We build upon previous work [12] where a corpus of German stock market reports between 1872 to 1930 was compiled for analyzing the sentiment over time. Sentiment analysis of the corpus aims to provide insight into the mood and opinions about the stock market during that period. While it is useful to consider the sentiment of an article or sentence in general as the aggregated sentiment of all statements, sentiment can also be expressed about specific aspects or entities. In the case of the stock market corpus we consider three levels of interest:

- **Individual Entities:** Sentiment towards specific entities of stocks that may be subject to a particular sentiment on a given day.
- **Sectors:** Statements towards sectors of the markets or groups of specific stocks, e.g. "the railway stocks were ...".
- **Overall:** The general mood at the stock market without specifying specific stocks or sectors.

The distinction between these different levels of sentiment analysis is crucial, since the historic texts tend to specifically emphasize opposing market movements, as can be seen in the following example:

> *Construction values dull throughout, only Deutsche Eisenbahnbau and Lindenbauverein again a little higher.*[1]

The example expresses a negative sentiment towards the construction value market, but highlights specific stocks (Deutsche Eisenbahnbau and Lindenbauverein) that traded higher. This type of sentiment analysis provides a more nuanced understanding of the sentiment towards the stock market during that time period. We regard these entity levels as the aspects of the aspect-based sentiment analysis.

B. Workflow and Data

To get a detailed understanding of the sentiment of the German historic stock market, we follow a three-step process: First we train a binary text classification model to identify if a sentence contains any sentiment at all, to filter out factual statements containing no sentiment. Second, we train a multi-label text classification model to detect which of the three levels are targeted by the expressed sentiment. Finally, we use the results of the entity-level classification to train an aspect-based sentiment model to extract the sentiment specifically with regard to the entity. This 3-step process is visualized in the left branch in Fig. 1. This enables us to analyze three sentiment time series with regard to the entity levels and also over all, if averaged.

To create a data set, a subsample of this corpus was annotated by an expert, as described in [13], and serves as

[1] Translated from German: *Bauwerthe durchweg matt, nur Deutsche Eisenbahnbau und Lindenbauverein wieder eine Kleinigkeit höher.*

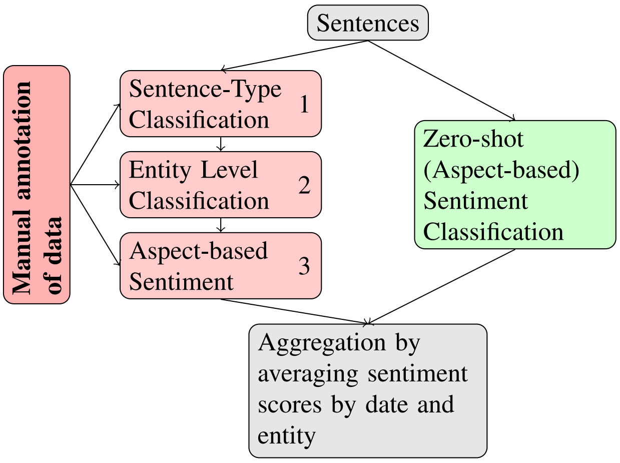
Sentences
Manual annotation of data
Sentence-Type Classification 1
Entity Level Classification 2
Aspect-based Sentiment 3
Zero-shot (Aspect-based) Sentiment Classification
Aggregation by averaging sentiment scores by date and entity

Fig. 1. Schematic drawing of the fine-tuned pipeline (left branch) and in the zero-shot pipeline (right branch). Red color indicates manual labelling or computational effort (training a modela).

training data. The data was sampled stratified over time to ensure that linguistic changes over time are represented in the data set. This results in three views on the data set:

- For sentence type classification there are 1651 examples, 609 neutral and 1042 containing a sentiment related statement.
- For sentences that contain any sentiment there are 732 sentence with at least one entity category assigned and a label density of 1.15.
- For aspect-based sentiment classification there are 1584 (sentence, aspect) pairs with an assigned sentiment of “positive", “negative" or “neutral".

Note that in our annotation scheme, “neutral" also includes calm or mixed statements, i.e. statements that have multiple contrary sentiments about an entity level or valuate it not in a positive or negative way. To simplify this into a common naming scheme we will refer to all of these as neutral, but it will be reflected in the hypothesis template of the zero-shot classification pipeline.

Using these three data sets, we build a fine-tuned pipeline of three models as shown in red in Fig. 1 that serves as a proxy of the expert solution to the task and will be the baseline to which we compare the zero-shot algorithm (shown in green).

C. Fine-tuned Pipeline

In this section we describe the fine-tuned pipeline, which consists of three separate models trained on one of the tasks corresponding to the left branch in Fig. 1. We only show a quick summary of the results that are discussed in Borst, Janos, Wehrheim, Lino, and Burghardt, Manuel [13]. As basis for every model, we use a German BERT variant pre-trained by the DBMDZ[2]. To evaluate the model, we split the annotated data into 80-20 training and validation splits, reported results are measured on the validation split. All three models use an Adam optimizer and after training the epoch with the best metric for the task is chosen.

TABLE I
RESULTS ON VALIDATION SET FOR THE ASPECT CLASSIFICATION STEP.

%	individual entities	sectors	market	micro avg	macro avg
precision	92.54	76.19	77.08	82.58	81.94
recall	89.86	84.21	90.24	88.02	88.11
f1-score	91.18	80.00	83.15	85.22	84.77

For sentence-type classification the transformer was fine-tuned as a binary classification model to distinguish neutral sentences from sentences containing sentiment statements, achieving 93% accuracy. The model was chosen because of the highest recall for sentences containing sentiment (96.5%). This ensures that we find most of the sentences containing sentiment statements.

Classifying which aspects the sentences contains was tackled as a multi-label classification problem with above-mentioned entity levels “individual entities", “sectors" and “market" as labels. The best model was chosen by macro average F1. The full results of this step is shown in Table Tab. I. We see quite balanced performance across all classes, with higher performance on individual entities. Individual entities have a very common linguistic pattern which makes them easy to detect.

In the third and final step we use the entity-level classification of step two as aspects and classify the combination of a (sentence, aspect)-pair into the sentiment classes “negative", “neutral" and “positive". A sentence can have multiple aspect-based sentiment annotations based on the result of the previous step. This model is trained as a single label classification task, that is, for every (sentence, aspect)-pair only one sentiment can be assigned. The best model was chosen by macro average F1 and achieves 80.7% accuracy and 80.7% macro F1.

D. Zero-Shot Pipeline

In this section we describe the pipeline to accomplish the same task without finetuning or training any model, corresponding to the right branch in Fig. 1. The aim is to perform the complex aspect-based sentiment classification process, described above, without using any of the knowledge that results from the manual coding and model training. This is especially important for aspect-based sentiment analysis, as we cannot assume knowledge about the type of entity contained in a sentence. We bypass this by classifying for the sentiments of all three entity categories and assume that, if there is no sentiment regarding any level this will result in a “neutral" label and will have no influence on the further analysis. We also classify the corpus with a zero-shot model with regard to overall sentence sentiment.

As zero-shot model, we use textual entailment classification, following the task description proposed in Yin, Hay, and Roth

[2] https://huggingface.co/dbmdz/bert-base-german-cased

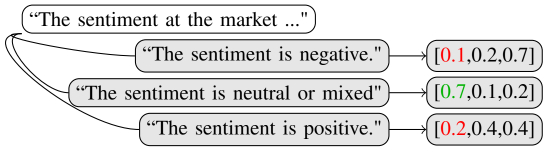
“The sentiment at the market ..."
“The sentiment is negative."
[0.1,0.2,0.7]
“The sentiment is neutral or mixed"
[0.7,0.1,0.2]
“The sentiment is positive."
[0.2,0.4,0.4]

Fig. 2. Schematic example for the formulation of the entailment task and its application to zero-shot text classification. The scores are the output for every sentence pair with regard to the categories *entailment*, *contradiction* and *neutral*. The highlighted numbers in color show the values that are compared with each other, which in this case would lead us to assign the category “neutral".

TABLE II
ZERO-SHOT EVALUATION METRICS ON THE MANUALLY LABELLED ASPECT-BASED DATA SET.

%	negativ	neutral	positiv	micro avg	macro avg
precision	75.69	60.48	81.27	67.61	72.48
recall	28.60	89.82	76.43	67.61	64.95
f1-score	41.52	72.29	78.77	67.61	64.19

[7] using a pretrained model from the huggingface hub[3]. 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 containing the label we want to classify. These hypotheses are created using a hypothesis template: *"The sentiment is [blank]"*[4]. The blank is then filled with the sentiment categories.

The model output provides a probability score for every premise and hypothesis pair and entailment class. We select the hypothesis with the highest probability of entailment as the classification result and assign the corresponding category. This leads to the formulation as show in Fig. 2. This approach is used for zero-shot sentiment classification.

For aspect-based zero-shot sentiment classification this approach can be extended by another placeholder in the hypothesis template, which is used to create the hypotheses. We use the template: *The sentiment for [aspect] is [label]*[5]. For every entity category above, we create the premise and hypothesis pairs by combining the entity category with each of the sentiment labels. Within each entity-level the procedure is the same as above. Fig. 3 shows a schematic drawing for this. The result of this step is an assignment of one sentiment label for every sentence and entity-level pair.

IV. EXPERIMENTS

Code and Data to replicate these findings can be found at https://git.informatik.uni-leipzig.de/computational-humanities/research/fedcsis-zero-shot-sentiment/

A. *Quantitative Comparison*

TABLE III
EVALUATION OF THE FINE-TUNED PIPELINE ON THE VALIDATION SET OF THE MANUALLY LABELLED ASPECT-BASED DATA SET.

	negativ	neutral	positiv	micro avg	macro avg
precision	81.6	90.5	85.7	86.1	85.9
recall	88.6	85.7	83.5	86.1	85.9
f1-score	84.9	88.0	84.6	86.1	85.9

TABLE IV
TABLE OF AGREEMENT BETWEEN THE ZERO-SHOT AND TRAINED PIPELINE ON THE ENTIRE CORPUS.

%	truth	negative	neutral	positive
zero-shot	negative	75.69	16.57	7.73
	neutral	30.54	60.48	8.97
	positive	9.49	9.25	81.27
trained	negative	84.76	11.43	3.81
	neutral	7.52	89.47	3.01
	positive	7.59	12.66	79.75

TABLE V
CONFUSION MATRICES OF THE TWO PIPELINES ON THE MANUALLY CODED VALIDATION SET.

	fine-tuned	negative	neutral	positive
aspect	zero-shot			
market	negative	85.79	6.79	07.42
	neutral	50.14	30.74	19.12
	positive	10.57	10.19	79.24
sectors	negative	83.86	8.76	7.37
	neutral	31.02	49.95	19.03
	positive	3.32	09.86	86.81
individual entities	negative	79.68	12.85	07.47
	neutral	33.73	42.23	24.05
	positive	2.91	09.54	87.55

1) *Data set metrics:* We evaluate the zero-shot algorithm on the same data used to train the fine-tuned pipeline on. Tab. II and Tab. III show the evaluation metrics for training and zero-shot respectively. Although there is a significant improvement in the F1-score of the trained model over the zero-shot model, it is noteworthy that this gap largely stems from the fact that the recall of negative sentiments is rather low. The precision for “negative" sentiments and all metrics for “positive" values are higher but a bit short of competing with the fine-tuned pipeline.

With further analysis, we find that the confusion matrices confirm the problem: Around 30% of predicted neutral labels are actually negative labels. This error is systematic, thus it may lead to an over-estimation of absolute values in the aggregated time series, but should not affect the overall trends.

2) *Agreement:* Tab. V shows the confusion matrix of the zero-shot pipeline and the fine-tuned pipeline. With regard to the manually coded data set, both algorithms seem to have comparable performance with strengths in classifying positive and negative examples. The confusion between neutral and

[3] https://huggingface.co/svalabs/gbert-large-zeroshot-nli
[4] Translated from German: *“Die Stimmung ist [label]."*
[5] Translated from German: *“Die Stimmung für [aspect] ist [blank]"*

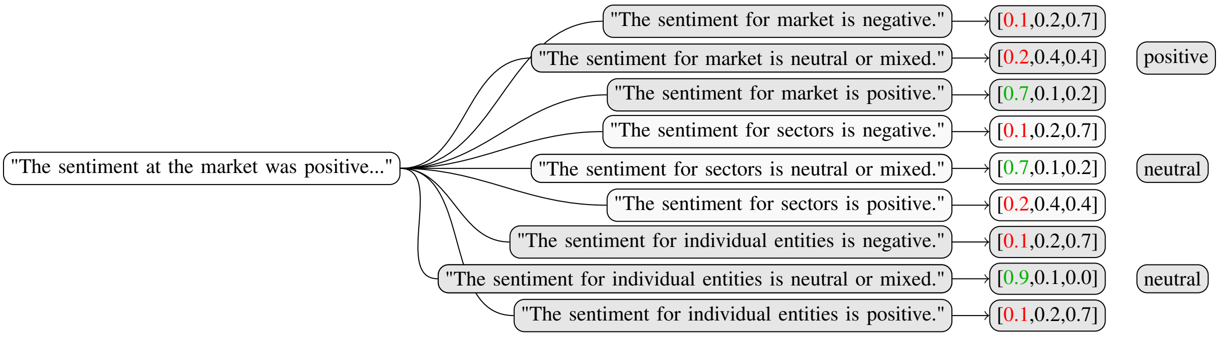
"The sentiment at the market was positive..."
"The sentiment for market is negative."
[0.1,0.2,0.7]
"The sentiment for market is neutral or mixed."
[0.2,0.4,0.4]
positive
"The sentiment for market is positive."
[0.7,0.1,0.2]
"The sentiment for sectors is negative."
[0.1,0.2,0.7]
"The sentiment for sectors is neutral or mixed."
[0.7,0.1,0.2]
neutral
"The sentiment for sectors is positive."
[0.2,0.4,0.4]
"The sentiment for individual entities is negative."
[0.1,0.2,0.7]
"The sentiment for individual entities is neutral or mixed."
[0.9,0.1,0.0]
neutral
"The sentiment for individual entities is positive."
[0.1,0.2,0.7]

Fig. 3. Schematic example for the formulation of the entailment task and its application to zero-shot text classification. The scores are the output for every sentence pair with regard to the categories *entailment*, *contradiction* and *neutral*. The highlighted numbers in color show the values that are compared with each other, which in this case would lead us to assign the category “neutral".

negative labels is the only position in the confusion matrix that seems to have substantially benefited from fine-tuning at all. All other label pairs show similar performance.

In Tab. IV we look at the confusion of agreement on entity level. In contrast to the previous table, high values now indicate that both pipelines agree to the same assignment of labels for any given (sentence, aspect) example. The table shows these statistics for the entire corpus (not only the validation data set). A very similar picture emerges: Both pipeline have a very high agreement about positive and negative examples for all entity levels. But there is considerable confusion if the zero-shot pipeline predicts neutral. For all entity levels there is a significant tendency that the fine-tuned pipeline would hand out a “negative" label where the zero-shot pipeline assigns “neutral". For the “market" level the agreement is even lower than 50%.

For the entity levels “sectors" and “individual entities" and on a global level, these errors have a systematic character that will not influence the overall trends considering that “most" prediction will still be correct.

3) Time Series Metrics: Besides an assessment of the quality of classification models, we want to compare resulting insights and possible analyses of both pipelines. To be able to use zero-shot instead of standard fine-tuning in a real-world application scenario, it should produce similar if not the same analysis result as the fine-tuned pipeline. In our case the basis of analysis are the sentiment time series that emerge from both of the pipelines. Comparing the time series expands the quantitative assessment of the classifier with the aspect of time. So the question we want to investigate here is: Would these pipelines create the same insights into the data?

The time series are generated by grouping the data over seven days and by summing up the sentiment labels. We evaluate negative, neutral and positive as “-1",“0", and “1" respectively. After that, time series are created by computing the rolling average over half a year (26 weeks). Time series are created for every entity level separately and for the overall sentiment. For overall sentiment we averaged the trained pipeline’s output per sentence to create one score and then did the same as above. Since we are not interested in absolute values, and we are dealing with a systematic error of the negative values, we also normalize each time series by mean-normalization. This normalization has no influence on trends or on the correlation factor, which we calculate below.

For comparison, we also applied an off-the-shelf neural network sentiment classification network for German language model for overall sentiment. Guhr et al. [32] train a “general-purpose German sentiment classification model". Since this model is off-the-shelf, it is not adapted to *historic* German language. However, we regard the comparison still as useful, since the availability of specific models is still one of the most common problems when researching textual data.

In the following, we want to evaluate the time series qualitatively and quantitatively. Fig. 4 shows a visual comparison of all these time series and will be discussed in the next section. For quantitative evaluation, we calculate the Pearson correlation of the zero-shot time series to the time series of the fine-tuned pipeline to verify if they have similar tendencies and trends. Tab. VI shows the Pearson correlation factors.

Fig. 4 reveals that the zero-shot and respective trained time series are often very close. In all cases except “market” the trends and tendencies are highly correlated. For “market” there is a period from 1900-1920 with a higher deviation. This leads to a Pearson correlation factor of only around 0.3. One explanation for this lies in the fact that while the concepts of “sector” and “individual entities” manifest in explicit syntactic figures and tokens, the concept of an “market” sentiment is not as easy to grasp sometimes. Another explanation lies in the fact that almost 50% of all entity mentions regard specific stocks or sectors. This might have higher influence on overall sentiment when aggregating over time. This argument is backed by the fact that the time series for individual entities shows the highest correlation of all time series. We also argue that it is not due to linguistic changes because these should affect all entity levels similarly in the same time periods, which we do not observe.

Additionally, we see that an off-the-shelve neural network sentiment classifier does not agree with the results of either zero-shot or fine-tuned pipeline. There is virtually no correlation of Guhr et al. [32] with the fine-tuned pipeline.

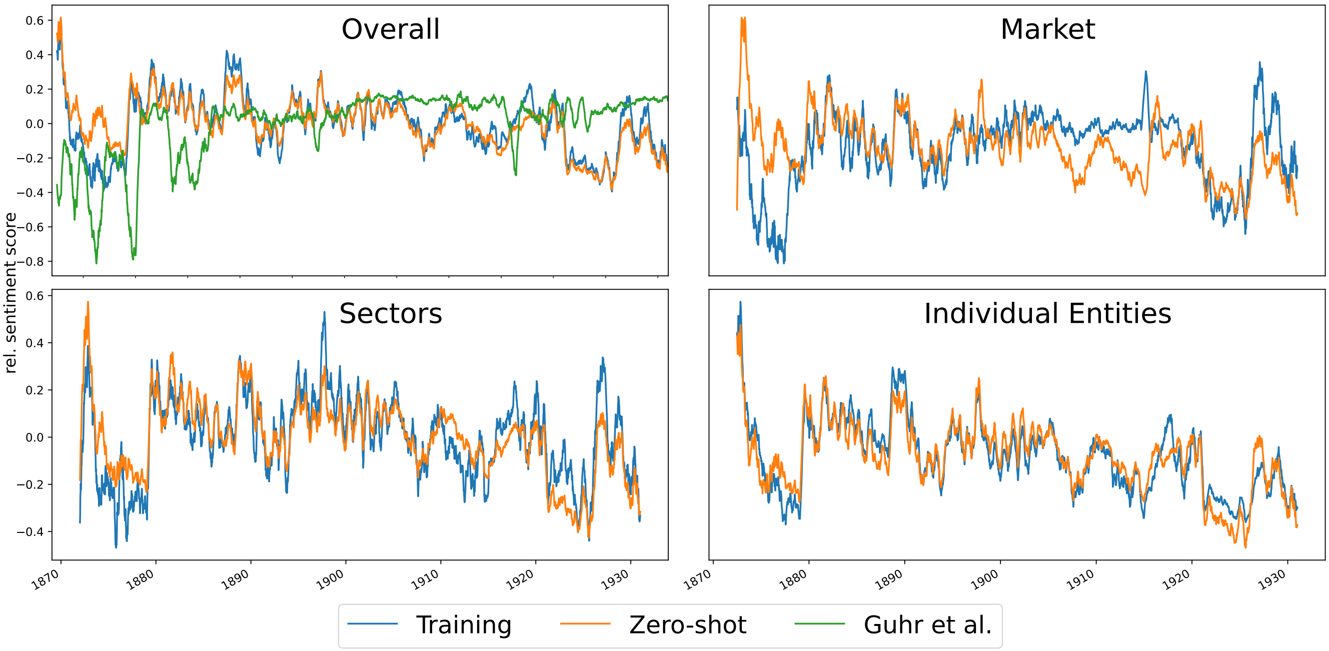

Fig. 4. Plot of the time sentiment time series for overall sentiments (top left) and for every entity level.

TABLE VI
PEARSON CORRELATION COEFFICIENT BETWEEN ZERO-SHOT AND TRAINED SENTIMENT TIME SERIES.

market	0.358472
sectors	0.822404
individual entities	0.911497
overall	0.870379
Guhr et al. [32]	0.077549

B. *Qualitative Assessment*

There has been some criticism about the application of entailment-based zero-shot recently [34]. The paper mentioned spurious correlation as a main driver of zero-shot classification performance in entailment-based solutions.

In our case this would be quite critical, because of formalisation of syntax and language in the historic realm could lead to considerable bias. We address this with a qualitative check of examples, on which the zero-shot and the trained pipeline disagree to identify patterns of errors.

As shown in Tab. IV, most of the disagreement occurs between neighbouring classes: positive and neutral , or negative and neutral, which in itself is often quite ambiguous. The only pattern we could find is that if there are opposing sentiments in one sentence such that the example should be labeled “neutral”, both algorithms seem to randomly pick one of the sentiments as the predicted polarity. This is not a systematic error but rather a random choice made by both pipelines. For instance, in the following example, two opposing polarities refer to the same entity level (“Individual Entities”). The trained pipeline assigns a positive polarity, whereas the zero-shot pipeline predicts it as neutral.

> *’Among foreign currency, Dutch lay firm, ruble notes continued to decline’.* [6]

In order to identify examples of linguistic patterns that are more difficult to classify, we examined instances where the zero-shot pipeline and trained pipeline assigned opposite polarities. One pattern that emerged with slightly higher frequency was related to the interpretation of the terms “supply” and “demand”[7]. While in general language, “supply” might have a positive connotation, in stock market reports, a predominance of supply is often associated with a high amount of selling and thus dropping prices, which is negatively connoted in this domain.

> *For Hansa shares, supply predominated.*[8]

> *Strongly in demand without supply were the 4% Reich and government bonds.*[9]

In the first example, the report mentions more supply than demand, which refers to falling prices. In this case, the trained pipeline correctly predicts a negative polarity while the zero-shot pipeline assigns a positive polarity. In the second example, the same situation occurs with “demand”, where this time the zero-shot pipeline correctly predicts a positive polarity while the trained pipeline assigns negative. Overall, the zero-shot pipeline tends to classify more sentences containing “supply”

[6] Translated from: *’Unter den fremden Devisen lagen holländische fest, Rubelnoten wurden weiter rückgängig.’*

[7] Translated from: *Angebot und Nachfrage*

[8] Translated from: *’Für Hansa-Aktien überwog Angebot.’*

[9] Translated from: *’Stark gefragt ohne Angebot wurden die 4% Reichs- und Staatsanleihen.’*

or “demand” as “neutral” (73%) than the trained pipeline (49%), which is in line with the above mentioned evaluation metrics.

While these examples highlight linguistic difficulties regarding the textual domain of historic stock market reports, there are no clear-cut patterns where one of the algorithms significantly fails to conform to our defined label scheme.

V. Practical and Epistemological Considerations

An alternative way to frame this paper's research question could be whether the zero-shot pipeline's results align with those based on human annotation, or more specific, how well the presumed annotation scheme conform with the zero-shot's definition of the annotation scheme. In our case, the results are generally encouraging thus far, but they also raise practical and epistemological follow-up questions, which we will outline in this section.

The comparison between the trained and zero-shot pipelines reveals that the assessment of the latter depends on the research question at hand. For researchers interested in sector- or entity-level sentiment, such as business historians, the zero-shot approach appears to be feasible, as evidenced by the evaluation metrics presented above. However, if one is interested in the market level, the differences between the two approaches appear to be too substantial, especially for the 1870s and 1910s. There are even differences in the degree of agreement with regards to different research objects within the same task, domain and time period.

Although we can only speculate about the reasons why the zero-shot approach does not agree with a model trained on human annotations for these periods, the lower performance at the market level, in general, is not surprising. Sentences referring to the “market” sentiment are more difficult to detect and interpret, even for human annotators, because the entity of “market” is often only mentioned implicitly. Ambiguity is a general problem in sentiment analysis. Furthermore, the fact that the zero-shot approach suffers from systematic biases may not be a problem if one is interested in time trends, but it may pose a problem in other cases. This is also true for other approaches, that can be used without training evaluation, e. g. dictionary-based methods.

Apart from these more technical aspects of our specific set-up, there are some epistemological reflections to be made. Researchers who consider using zero-shot methods because they lack the resources to create training data and fine-tuning a model, face a fundamental dilemma - as anyone using unsupervised tools does: How can we rely on the results if we do not have data for a formal evaluation and if we have data to evaluate why not also fine-tune? When looking at the time series in Fig. 4, even the most well-versed domain expert may struggle to discern whether these results are mere artifacts created by an algorithm or substantial results. Furthermore, even if the expert can discern the results, what would we learn from these results that we did not know before? In any case there is the possibility of confirmation bias, when not properly supported by close reading. In other words, we face the paradox that the very advantage of zero-shot models is also a considerable drawback for practical application.

Of course, there is general evidence for the quality and performance of zero-shot models, where especially polarity classification has been shown to work more consistently. But there is no general notion of why this should be transferable to another domain or another language with a “similar” task. The problem of distributional shift or domain adaptation often contributes to loss of performance [35, 36, 37]. Then again: How similar do these domains have to be, to safely assume generalization? These questions are particularly challenging to answer for historiographic research, which often covers very specific domains and languages or longer time periods for which there are hardly any pre-checked settings to be found.

Finally, this study raises a lot of questions regarding implicit assumptions when applying zero-shot sentiment classification, like: Is there a “correct" sentiment in these texts? If so, are expert-level humans able to identify it, correctly reflecting the historic reality? Does the technical ability of these models to generalize suffice in this scenario?

Recent research suggest that every step along the way to a trained pipeline based on human annotations involves the risk of bias [38]. Also, in the special case of sentiment, there are many studies that evidence that scholar's and crowd-sourced annotations alike have particularly low agreement between annotators in historic texts [39, 36, 37]. Even the agreement of various machine-learning algorithms seems surprisingly low even when trained and evaluated on the same sentiment data sets [40]. There is much work done in the field of domain adaptation in sentiment classification [35, 41, 42, 43], but all these implicitly rely on the human annotations as performance metric as well. All these factors contribute to the uncertainty of “correct" results in any case, but might also lead to the conclusion that a zero-shot approach may suffice to discover underlying trends.

There is no space here to provide answers to these questions, nor do we claim to have any. In the end, the usefulness of zero-shot learning will probably depend on the research question, the domain, the possibility to conduct some sort of evaluation (e.g. at least some human annotations), and maybe the general willingness to trust unsupervised approaches, which is distributed unevenly across different research communities. These issues will, however, get more pressing with the wider availability of powerful zero-shot tools like GPT-4.

VI. Conclusion

In this study, a zero-shot text classification pipeline was applied to an aspect-based sentiment analysis of German historic texts of stock market reports in the digital humanities. The goal was to get insights in how useful these methods are in a real application scenario. We provided in-depth comparisons, qualitatively and quantitatively, between a fine-tuned pipeline and a zero-shot pipeline. The results show that both can deliver usable results in our aspect-based sentiment analysis. The trends and insights produced by the zero-shot models were highly correlated with those produced by the trained models. They were found to be particularly useful for classical sentence polarity classification, but also performed well for aspect-based sentiment analysis. Even if the results may differ in some details and there are systematic errors, we can confidently say that zero-shot models provide a good exploration tool and an easy start for sentiment analysis, especially in cases where no hand-coded data exists.

The zero-shot models were also found to work better than an off-the-shelf BERT German sentiment model. Since the general availability of domain-specific models is still not fully achieved, especially for languages other than English, zero-shot approaches to sentiment analysis may help to close the gap. Also, zero-shot models for aspect-based sentiment were found to work better with concrete entities (target text) rather than general aspects, which we were not able to consider in this study. We defer this task to future work on the subject.

However, it should still be noted that correlation and data set metrics are just a hint at performance and that the factual correctness of the sentiment analysis is difficult to prove, as this would require manual examination of the textual content and sentiment analysis still contains a subjective nature. We only compared the zero-shot results to the result of the trained pipeline to answer the question, if both models would provide similar insights, while the interpretation of these results is not part of this paper.

Still, we believe that zero-shot models in sentiment analysis, and for text classification in general, seem a promising approach, especially when considering the progress and the potential of models similar to GPT-4.

Acknowledgements

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

References

[1] Bing Liu. *Sentiment Analysis and Opinion Mining*. Vol. 5. 2012.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[3] Tom Brown et al. "Language Models are Few-Shot Learners". en. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[4] Y. Xian, B. Schiele, and Z. Akata. "Zero-Shot Learning — The Good, the Bad and the Ugly". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017, pp. 3077–3086.

[5] Lei Shu et al. *Zero-Shot Aspect-Based Sentiment Analysis*. 2022. arXiv: 2202.01924 `[cs.CL]`.

[6] Kishaloy Halder et al. "Task-Aware Representation of Sentences for Generic Text Classification". en. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 3202–3213. DOI: 10.18653/v1/2020.coling-main.285.

[7] Wenpeng Yin, Jamaal Hay, and Dan 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*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 3912–3921. DOI: 10.18653/v1/D19-1404.

[8] Senait Gebremichael Tesfagergish, Jurgita Kapočiūtė-Dzikienė, and Robertas Damaševičius. "Zero-Shot Emotion Detection for Semi-Supervised Sentiment Analysis Using Sentence Transformers and Ensemble Learning". In: *Applied Sciences* 12.17 (2022), p. 8662. DOI: 10.3390/app12178662.

[9] Mengting Hu et al. "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)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021, pp. 6330–6340. DOI: 10.18653/v1/2021.acl-long.495.

[10] Ronald Seoh et al. "Open Aspect Target Sentiment Classification with Natural Language Prompts". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 6311–6322. DOI: 10.18653/v1/2021.emnlp-main.509.

[11] Anindya Sarkar, Sujeeth Reddy, and Raghu Sesha 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. Tokushima, Japan: Association for Computing Machinery, 2019, pp. 49–56. DOI: 10.1145/3342827.3342850.

[12] Wehrheim, Lino et al. "„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*. Dhd23 Open Humanities Open Culture. Trier, 2023. DOI: 10.5281/zenodo.7688632.

[13] Borst, Janos, Wehrheim, Lino, and Burghardt, Manuel. ""Money Can't Buy Love?" Creating a Historical Sentiment Index for the Berlin Stock Exchange, 1872–1930". In: *Book of Abstracts*. Digital Humanities. Graz, 2023.

[14] Evgeny Kim and Roman Klinger. “A survey on sentiment and emotion analysis for computational literary studies”. In: *Zeitschrift für digitale Geisteswissenschaften* (Aug. 2019). DOI: 10.17175/2019_008_v2.

[15] George Akerlof and Robert Shiller. *Animal Spirits: How Human Psychology Drives the Economy and Why It Matters for Global Capitalism*. Vol. 21. Jan. 1, 2009. ISBN: 978-0-691-14592-1. DOI: 10.2307/j.ctv36mk90z.

[16] Paul C. Tetlock. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market”. en. In: *The Journal of Finance* 62.3 (2007), pp. 1139–1168. DOI: 10.2139/ssrn.685145.

[17] Diego García. “Sentiment during Recessions”. en. In: *The Journal of Finance* 68.3 (2013), pp. 1267–1300. DOI: 10.1111/jofi.12027.

[18] Alan J. Hanna, John D. Turner, and Clive B. Walker. “News media and investor sentiment during bull and bear markets”. en. In: *The European Journal of Finance* 26.14 (Sept. 2020), pp. 1377–1395.

[19] Kostadin Mishev et al. “Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers”. In: *IEEE Access* 8 (2020). ISSN: 2169-3536.

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

[21] Zhuang Liu et al. “FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI’20. Yokohama, Yokohama, Japan, 2021. DOI: 10.24963/ijcai.2020/615.

[22] Pekka Malo et al. “Good debt or bad debt: Detecting semantic orientations in economic texts”. In: *Journal of the Association for Information Science and Technology* 65.4 (Apr. 2014), pp. 782–796.

[23] Ankur Sinha et al. “SEntFiN 1.0: Entity-aware sentiment analysis for financial news”. In: *Journal of the Association for Information Science & Technology* 73.9 (2022), pp. 1314–1335.

[24] Wenxuan Zhang et al. “A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges”. In: *IEEE Transactions on Knowledge and Data Engineering* (2022). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 1–20. ISSN: 1558-2191. DOI: 10.1109/TKDE.2022.3230975.

[25] Thomas Wolf et al. “HuggingFace’s Transformers: State-of-the-art Natural Language Processing”. In: *Computing Resource Repository* abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771.

[26] Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 2018, pp. 328–339. DOI: 10.18653/v1/P18-1031.

[27] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 2019, pp. 5754–5764.

[28] Jonathan Bragg et al. “FLEX: Unifying Evaluation for Few-Shot NLP”. In: *Neural Information Processing Systems*. 2021. DOI: 10.1162/tacl_a_00485.

[29] Yujia Bao et al. “Few-shot Text Classification with Distributional Signatures”. In: *International Conference on Learning Representations*. 2020. DOI: 10.1145/3531536.3532949.

[30] Yaqing Wang et al. “Generalizing from a Few Examples: A Survey on Few-Shot Learning”. In: *ACM Comput. Surv.* 53.3 (June 2020). DOI: 10.1145/3386252.

[31] Edgar Schonfeld et al. “Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders”. en. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 8239–8247. DOI: 10.1007/s00521-022-07413-z.

[32] Oliver Guhr et al. “Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems”. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1627–1632. ISBN: 979-10-95546-34-4.

[33] Francesco De Toni et al. “Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0”. In: *ArXiv* abs/2204.05211 (2022). DOI: 10.18653/v1/2022.bigscience-1.7.

[34] Tingting Ma et al. “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)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, Aug. 2021, pp. 786–796.

[35] Chenggong Gong, Jianfei Yu, and Rui Xia. “Unified Feature and Instance Based Domain Adaptation for Aspect-Based Sentiment Analysis”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Association for Computational Linguistics, Nov. 2020, pp. 7035–7045. DOI: 10.18653/v1/2020.emnlp-main.572.

[36] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. “Emotions from Text: Machine Learning for Text-based Emotion Prediction”. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. HLT-

EMNLP 2005. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 579–586. DOI: 10.3115/1220575.1220648.

[37] Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. *Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior*. Aug. 1, 2018.

[38] Mihir Parmar et al. “Don’t Blame the Annotator: Bias Already Starts in the Annotation Instructions”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2023. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1779–1789.

[39] Rachele Sprugnoli et al. “Towards sentiment analysis for historical texts”. In: *Digital Scholarship in the Humanities* 31.4 (July 2015), pp. 762–772. DOI: 10.1093/llc/fqv027.

[40] Frank Xing et al. “Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. COLING 2020. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 978–987. DOI: 10.18653/v1/2020.coling-main.85.

[41] Mohammad Rostami and Aram Galstyan. *Domain Adaptation for Sentiment Analysis Using Increased Intraclass Separation*. July 4, 2021. arXiv: 2107.01598[cs]. URL: http://arxiv.org/abs/2107.01598 (visited on 05/22/2023).

[42] Guoliang Kang et al. “Contrastive Adaptation Network for Unsupervised Domain Adaptation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019, pp. 4888–4897.

[43] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. “Projecting Embeddings for Domain Adaption: Joint Modeling of Sentiment Analysis in Diverse Domains”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 818–830.