A Dataset for Explainable Sentiment Analysis in the German Automotive Industry

Andrea Zielinski Calvin Spolwind
Anna Grimm Henning Kroll
Fraunhofer ISI
Breslauer Str. 48
76139 Karlsruhe
andrea.zielinski@isi.fraunhofer.de

Abstract

While deep learning models have greatly improved the performance of many tasks related to sentiment analysis and classification, they are often criticized for being untrustworthy due to their black-box nature. As a result, numerous explainability techniques have been proposed to better understand the model predictions and to improve the deep learning models. In this work, we introduce InfoBarometer, the first benchmark for examining interpretable methods related to sentiment analysis in the German automotive sector based on online news. Each news article in our dataset is annotated with respect to the overall sentiment (i.e., *positive, negative and neutral*), the target of the sentiment (focusing on *innovation-related topics such as e.g. electromobility*) and the rationales, i.e., textual explanations for the sentiment label that can be leveraged during both training and evaluation. For this research, we compare different state-of-the-art approaches to perform sentiment analysis and observe that even models that perform very well in classification do not score high on explainability metrics like model plausibility and faithfulness. We calculated the polarity scores for the best method BERT and got a macro F1-score of 73.8. Moreover, we evaluated different interpretability algorithms (LIME, SHAP, Integrated Gradients, Saliency) based on explicitly marked rationales by human annotators quantitatively and qualitatively. Our experiments demonstrate that the textual explanations often do not agree with human interpretations, and rarely help to justify the models decision. However, global features provide useful insights to help uncover spurious features in the model and biases within the dataset. We intend to make our dataset public for other researchers.

1 Introduction

There is a wealth of economic (online) news so that even specialists find it challenging to analyze all information in a timely manner. Therefore, methods that seek to automatically identify the sentiment in textual data for calculating economic indices or trends have become prominent in recent years (Seki et al., 2022; Kalamara et al., 2022; Levenberg et al., 2014; Weiss and Nemeczek, 2022).

In particular, deep learning based approaches have yielded high performance (Poria et al., 2020; Rojas-Barahona, 2016; Hartmann et al., 2022), and their results even seem to correlate with survey-based indicators (Marszal et al., 2022; Katayama et al., 2019).

However, sentiment analysis[1] is challenging due to the domain-specific language and unavailability of labeled datasets needed for training or fine-tuning neural models (Araci, 2019; Mishev et al., 2020). It is also a complex task, as a document may contain multiple targets whose sentiments may be opposite. For example, in the sentence *Um gute Produkte und Leistungen zu liefern, hat die Firma hohe Investitionen getätigt. (In order to deliver good products and services, the company has made high investments.)*, the author expresses a positive sentiment towards *products and services* using the opinion words *gute (good)* and a negative sentiment towards *Investitionen (investments)* using *hohe (high)*. Apart from mixed sentiment, another challenge is that depending on the context, the sentiment may also shift to the opposite, e.g., *hoher Komfort/hoher Verlust (high comfort vs. high loss)*.

Since model interpretability is a critical requirement for many downstream applications, recently explainable artificial intelligence (XAI) techniques that make models explainable have been proposed (Danilevsky et al., 2020; Madsen et al., 2021), and also been adopted for sentiment analysis in general (El Zini et al., 2022) or with respect to financial news (Gite et al., 2021; Xing et al., 2020). In line with Arrieta et al. (2020), we consider that an AI

[1] In our work, we focus on document-level sentiment analysis, i.e. the goal to infer the overall opinion of a news article, which is assumed to convey a unique opinion towards a topic.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 138–148
July 14, 2023 ©2023 Association for Computational Linguistics

system is explainable, if the task model is intrinsically interpretable or if it is complemented with an interpretable and faithful explanation. In this work, we focus on transparent inference through the generation of a post-hoc explanation for the final output prediction (Thayaparan et al., 2020).

It is well known that machine learning techniques suffer from an inadequate generalization capacity on real-world datasets, which are often scarce and imbalanced. While transfer learning and regularization techniques have been proposed to solve this issue (Pan and Yang, 2010; Srivastava et al., 2014), overfitting the training data is still a common problem, especially for non-English texts. As has been stated by various authors, explainable modelling can contribute to trustworthy AI systems, that go beyond quantitative performance metrics such as accuracy (Danilevsky et al., 2020; Wallace et al., 2020).

To this aim, we build up a corpus of German news articles that report recent innovations in the automotive field, which includes class labels for topics such as *e-mobility, autonomous driving, Diesel, hydrogen, synthetic fuel, and misc*, sentiment labels *i.e., positive, neutral, negative*, and human rationales for studying trustworthiness, inspired by recent work in this field (Zaidan and Eisner, 2008; Wang et al., 2021, 2022). We conduct experiments building on state-of-the art machine learning models for sentiment analysis, as well as interpretation methods (Molnar, 2022) using established evaluation metrics for interpretability (DeYoung et al., 2020; Atanasova et al., 2020). Human explanation for sentiment classification is essential for high-impact settings such as financial analysis where human rationales are required to support specialists' decisions. Basic research questions are

- **RQ1** How well can we automatically classify news articles as expressing a *positive, negative or neutral* opinion at the document-level and which approach gives the best performance. How well is human IAA for our corpus.

- **RQ2** Which sentiment detection and XAI method delivers interpretable results, highlighting words or phrases that have lead to the sentiment class. Is this also reflected by human rationales on positive or negative aspects?

Our main contributions are as follows:

- We present a benchmark for rationalized predictions, including baseline models and results. We quantitatively compare SVM, CNN, LSTM and BERT models in the context of sentiment analysis by performing an evaluation on our manually annotated corpus. Using local and global explanation methods, we are able to retrieve the tokens that are most indicative of the sentiment class label.

2 InfoBarometer Dataset

In this section, we provide the dataset collection strategy, the annotation procedure and statistics on our dataset.

2.1 Data Collection

We scraped German news articles related to recent innovative trends in the automotive sector for the time period Feb-2022 to Dec-2022 from online news providers[2], filtering the results by selecting innovation-related keywords, and then applying a precision-oriented topic classifier.

We keep a minimal layout with title and paragraphs, eliminating links and pictures in the news articles.

2.2 Annotation Procedure

The dataset was annotated using *tagtog*[3], a web-based tool, which was configured appropriately for our annotation task. Each news article contains three types of annotations: the sentiment polarity, the topic category, and the human rationales.

Regarding **sentiment polarity**, each news article is assigned one of the following polarity labels: *positive, negative, neutral* which reflects the overall sentiment label about a certain **topic category** that is prevalent in a given news article, and chosen from a predefined inventory of categories, i.e. *autonomous driving*, *electromobility*, *hydrogen*, *Diesel*, *Synfuel* and *misc* (see Appendix A). We further ask the annotators to highlight **rationales** as text spans, that could justify the final polarity annotation.

Annotation guidelines have been set up that clearly explain the goal of the annotation task, how to annotate tokens or spans and also include a definition for each topic category, following best-practice recommendations (Wiegreffe and Marasović, 2021).

[2] www.automobil-industrie-vogel.de, www.automobilwoche.de

[3] https://www.tagtog.com/

We provided multiple examples with topic and polarity classification as well as rationale annotations to help the annotators understand the task.

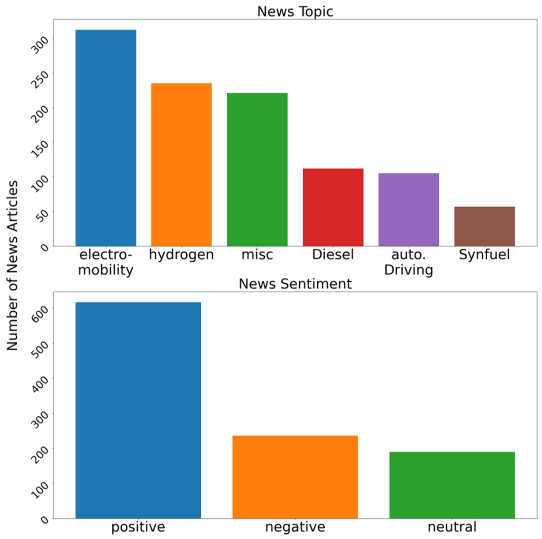


Figure 1: Dataset Statistics

Each news article was annotated by 3 annotators, experts in the automotive domain. In addition a master annotator consolidated annotations and decided on the final label and rationales, when annotators disagreed. Since classifying implicit sentiments is a challenging task, which demands expert knowledge and requires interpretation, we conducted an inter-annotator agreement study to determine whether our annotation guidelines were properly defined and resulted in consistent and reproducible annotations. To measure the interannotator agreement of the manual annotation (IAA), we calculate the overall F1-measures of the annotations, treating the master annotation as the gold standard, resulting in an average F1 score of 0.829 and 0.774 for text classification and sentiment annotation, respectively, which indicates a high agreement of the labeled data. The IAA is depicted in Table 1.

IAA	Topic Label f1	Sentiment f1
annotator1	0.8249	0.7989
annotator2	0.7716	0.6821
annotator3	0.8905	0.8396

Table 1: Inter-Annotator Agreement of the Sentiment and Topic Classification Task

2.3 Dataset Statistics

The dataset is composed of 1,039 news articles from the web. As can be seen in Figure 1, the extent to which German news media cover innovation-related technologies in the automotive sector differs a lot in terms of frequency of certain topics. Looking at the sentiment distribution, we see that they mostly receive positive coverage in German news media with 59.12%, while 22.65% are negative and 18.23% of all news is neutral. Given these results, we infer that news on innovation and new technologies are indeed overall positive for the German automotive industry. While sentiment polarity annotations express overall sentiments (positive/negative/neutral) in the news article, it might nonetheless include different opinionated expressions with different polarities, positive or negative, indicated by spans of rationales on each news item.

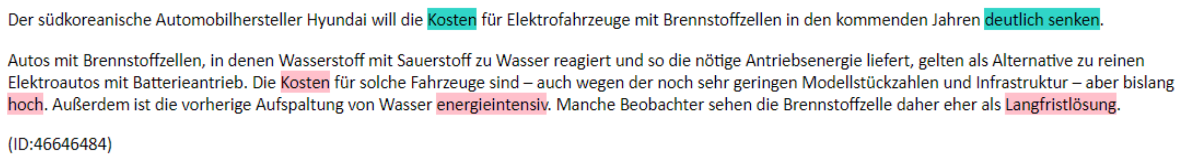
Figure 2 shows all rationales that have been selected for a specific news article, grounding the overall positive sentiment annotation for the whole article in the positive evidences *Steigerung (increase), gerettet (saved), mehr als verdoppelt (more than doubled)*. The aim was to annotate the news in an exhaustive way, which in this example includes also negative evidences (e.g. *Mangel an Halbleitern (lack of semiconductors), verfehlt (missed), Produktion gebremst (hampered production)*.

The InfoBarometer dataset has been split into training, validation, and testing set. The training set contains 833 news items, while the dev and test set contain 104 and 102 items, respectively. The average length for each news article is 226.21 tokens, consisting of 13.23 sentences on average.

For this dataset, the number of rationales per instance is 6.25 (9.97 tokens) on average, which varies slightly by sentiment class, also when taking the average percentage of rationales to input sequence length into account (see Table 2). Note that there is no perfect correlation between the length of an article and the length of the rationale based on our Benchmark Corpus (see Fig. 3). Only the test set was used in the plausibility evaluation.

2.4 Related Work

Many datasets with human-annotated rationales have been published for interpretability evaluation, in particular highlight-based rationales (DeYoung et al., 2020). For the task of sentiment analysis, the Movie Reviews Dataset (Zaidan et al., 2007) has lately been extended (DeYoung et al., 2020) to



Der südkoreanische Automobilhersteller Hyundai will die Kosten für Elektrofahrzeuge mit Brennstoffzellen in den kommenden Jahren deutlich senken.

Autos mit Brennstoffzellen, in denen Wasserstoff mit Sauerstoff zu Wasser reagiert und so die nötige Antriebsenergie liefert, gelten als Alternative zu reinen Elektroautos mit Batterieantrieb. Die Kosten für solche Fahrzeuge sind – auch wegen der noch sehr geringen Modellstückzahlen und Infrastruktur – aber bislang hoch. Außerdem ist die vorherige Aufspaltung von Wasser energieintensiv. Manche Beobachter sehen die Brennstoffzelle daher eher als Langfristlösung.

(ID:46646484)

Figure 2: Annotation of a German news article. Human Rationales are highlighted in blue (positive) and red (negative).

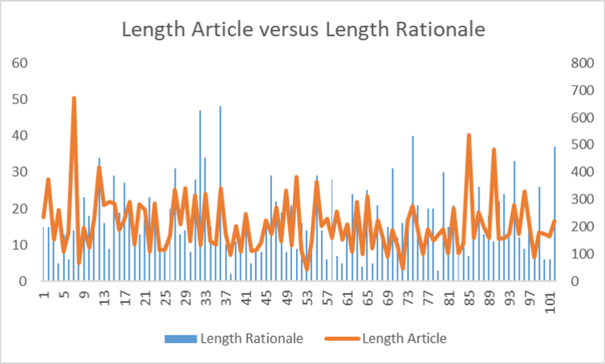


Figure 3: Dataset Statistics: The correlation between the length of an article and the length of the rationale based on our Benchmark Corpus.

Sentiment	Rationales	Len. Rationale	Len. Ratio
negative	6.75	16.46	7.27
neutral	5.52	14.90	6.59
positive	6.22	17.79	7.86

Table 2: The Rationales column presents the average number of rationales, the Length Rationale the average number of tokens per instance, while the Length Ratio column presents the average ratio of rationale to input sequence length measured in tokens.

build a comprehensive evaluation benchmark for interpretability (Zaidan and Eisner, 2008).

Wang et al. (2021) propose a novel evaluation dataset for sentence-level sentiment analysis for English. They provide highlight-based rationales to evaluate the robustness and interpretability (i.e., faithfulness and plausibility) of various algorithms (Lime, SHAP, attention) and deep learning frameworks, including LSTM and SKEP (Tian et al., 2020).

In comparison to these datasets, the InfoBarometer dataset also contains span-level rationale annotations, however, we annotated three polarity classes instead of 2, including the neutral class which either denotes the lack of sentiment towards a topic or may consist of an equal amount of positive and negative evidence in the news article. The included rationale annotations are comprehensive, since the annotators were asked to mark all text supporting the polarity label and we aggregated the rationales from multiple annotators if they agreed on the same class. In addition, approx. 20% of all article contain mixed sentiments with evidence for both the positive and negative polarity class.

3 Methodology

Due to the small size of our dataset, we think that there is high need for explainability, regarding model interpretability. Through our experiments, we like to investigate if explainability techniques can uncover biases within the dataset and check the trustworthiness of the ML models trained on the InfoBarometer dataset.

3.1 Classification Models

We consider three model families of text encoders with increasing complexity: a support vector machine (SVM) with a linear kernel (Pang et al., 2008), a CNN (Kim, 2014), LSTM (Hochreiter and Schmidhuber, 1997) and a pre-trained BERT model (Devlin et al., 2018)[4]. To build the sentiment classifier, we fine-tune BERT on our dataset to encode domain specific semantics and augment it with a prediction task to encode sentiment and topic information. In the CNN and LSTM models, the word inputs are initialized to 300-dimensional GloVe embeddings (Pennington et al., 2014)[5]. We tune all parameter using the validation set and use the best model for testing. While recently contextual language models like BERT performed best on a variety of benchmark datasets for document-sentiment analysis, also CNNs and RNNs have been applied successfully (Poria et al., 2020), as well as sequence models (Li et al., 2016), because they can encode long-range dependencies in the word sequence, which is fundamental to model

[4] German BERT model is from Huggingface https://huggingface.co/deepset/gbert-base

[5] GLOVE embeddings are from Deepset https://www.deepset.ai/german-word-embeddings

negation and polarity shifts. On the movie dataset[6] (Zaidan and Eisner, 2008), CNNs based on pre-trained embeddings showed highest performance in terms of accuracy, outperforming RNNs and LSTMs, because they are able to learn contextual semantic features that are relevant for the sentiment prediction task. Our baseline models are:

Convolutional Neural Networks: CNNs (Kim, 2014) are specially powerful in exploiting the local correlation and pattern of the data by their feature maps. Since parallelization is possible, they are more efficient than LSTMs and yield a good performance for classification tasks with little fine tuning (Nedjah et al., 2022).

Long-short time memory : LSTM (Schmidhuber et al., 1997) is a popular recurrent neural network architecture for modeling sequential data, and can capture long term dependencies. They have the characteristics of selectivity and memory cells and solve the gradient vanishing problem.

Bidirectional Encoder Representations from Transformers (BERT) by (Devlin et al., 2018), enriched with the deep bidirectional word representations released by HuggingFace (Wolf et al., 2019). Key aspects of the BERT model include multi-head self-attention as well as a Transformer layer.

See Appendix A.2 for our hyperparameter settings.

3.2 Explainability Models

We focus on post-hoc feature attribution explanation methods, probing the model parameters and/or input-output pairs of an already trained black-box model. We use Captum[7], an open source library for model interpretability built on PyTorch for our experiments, including explanation methods that are *gradient-based*, i.e. InputXGradient (Shrikumar et al., 2016), Saliency (Simonyan et al., 2013) *perturbation-based*, i.e. Shapley Value Sampling (Shapley, 1953; Castro et al., 2009), and Lime (Ribeiro et al., 2016) as a *model simplification* method. Saliency maps are used to visualize which parts of the input are responsible for the prediction. In the case of gradient-based methods, we measure the importance of a feature using the derivative with respect to that feature. We apply the L2 norm to obtain the magnitude of a gradient vector, which becomes a saliency of each token. As the output, we take the loss with the top prediction as the ground truth class (Han et al., 2020).

4 Experimental Setup

We execute experiments on topic classification and sentiment analysis for four different ML approaches. The model predictions as well as the XAI models' explanations are compared to ground truth data that has been annotated by domain experts. For an overview of the proposed approach see Figure 4.

4.1 Classification Results

Following prior works, we conduct experiments with all described sentiment classification models.

Sentiment Model	Macro F1	Accuracy
SVM_linear	0.5320	0.6476
CNN with GLOVE	0.3824	0.6095
LSTM with GLOVE	0.4013	0.5048
Fine-tuned BERT	0.7376	0.7809

Table 3: Accuracy of sentiment analysis models (in terms of classification accuracy and macro f1), evaluated on the news datasets with 1039 articles devided into train/dev/test.

Topic Classifier Model	Macro F1	Accuracy
SVM_linear	0.3600	0.4857
CNN with GLOVE	0.4981	0.6381
LSTM with GLOVE	0.3248	0.5143
Fine-tuned BERT	0.7904	0.8190

Table 4: Accuracy of the topic classification models (in terms of classification accuracy and macro f1), evaluated on the news datasets with 1039 articles devided into train/dev/test.

4.2 Computational Efficiency

We also compare the computational efficiency of our ML models and XAI techniques (see Section 5.3) that are critical in a setting which requires timely decision support. We recorded the computational time to generate sentiment and class predictions on a computer cluster with 2 AMD EPYC 7742 64-Core Processors 2.25 GHz, 192 GB RAM, x64 NVIDIA A100-PCIE 40GB.

The computational time for classifying our test dataset is shown in Table 8. The speed for testing is relatively low compared to the time for training the model, so that all of them can be used in a real-time interactive system.

[6] https://www.tensorflow.org/datasets/catalog/movie_rationales

[7] https://captum.ai/

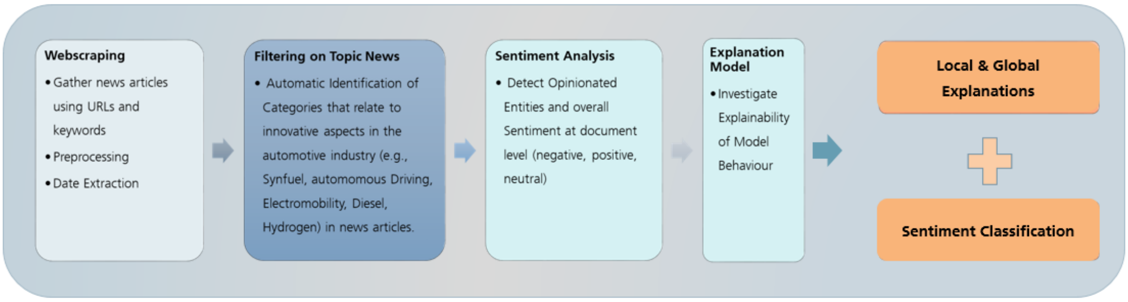

Webscraping
• Gather news articles using URLs and keywords
• Preprocessing
• Date Extraction
Filtering on Topic News
• Automatic Identification of Categories that relate to innovative aspects in the automotive industry (e.g., Synfuel, automomous Driving, Electromobility, Diesel, Hydrogen) in news articles.
Sentiment Analysis
• Detect Opinionated Entities and overall Sentiment at document level (negative, positive, neutral)
Explanation Model
• Investigate Explainability of Model Behaviour
Local & Global Explanations
Sentiment Classification

Figure 4: Overview of the proposed approach

Topic Classifier Model	Train sec	Test sec
CNN with GLOVE	3228.06	27.85
LSTM with GLOVE	703.44	22.83
Fine-tuned BERT	1062.33	21.54

Table 5: Computational efficiency. We report the mean speed in seconds for topic and sentiment classification, for training on train/dev and testing on the test dataset

4.3 Explainability Results

Global Explanation We use SHAP to compute the global features that influence the overall performance of our ML models (Lundberg and Lee, 2017). A key advantage of SHAP is that feature contributions are all expressed in terms of the outcome variable, providing a same scale to compare the importance of each feature against each other. Figure 5 shows what features are important to the model overall. Top salient features in our test set for the negative and positive class are *stop, decrease, bad, crises, expensive* and *preferred, solve, build up, possible, modern, funding*, respectively.

Local Explanation An example of a local explanation from our dataset is shown in Figure 6. We can see that the sentence has been predicted correctly to be positive. However, our explainability methods partly disagree on the interpretation of the same prediction made by the same BERT model. XAI methods depict either *investiert (invest), synthetisch (synthetic)* or *Bundesregierung (government)* as most salient feature that contribute to a positive sentiment assessment.

We observe that the model is relying on spurious features, like *electric* and *autonomous*. Furthermore, any bias in the data, e.g. mentions of nationalities like *Chinese*, need to be eliminated before deployment.

4.4 Evaluation of Explainability

4.4.1 Metrics

Plausibility: In the context of this work, we generate token-level explanations, selecting a list of the top k most salient tokens. For each instance, the model must generate an explanation defined as a subset of zero or more tokens from the instance. The longer the instance, the more explanation tokens are selected. We use IOU (Intersection-Over-Union) F1-score and Token F1 score to measure plausibility (DeYoung et al., 2020). We compute the score at the token level and do not consider continuous sub-strings (spans), since this metrics is considered too imprecise (Wang et al., 2022).

Faithfulness: For debugging, it is important to produce explanations that are faithful, i.e., accurately reflect the features considered important by the model (Jacovi and Goldberg, 2020). The XAI method should faithfully reveal information about the model's inner working. A common approach is to iteratively mask salient features from the input and measure the average drop in the model's performance (DeYoung et al., 2020). In this work, we follow the approach of Atanasova et al. (2020), which relies on producing several perturbations by masking [0, 10, 20,.., 100%] of the input tokens in order of decreasing saliency, and then computing the area under the threshold-performance.

5 Results - Performance Comparison for XAI Methods

We compare LIME, Saliency, InputXGradient and SHAP with respect to the evaluation criteria plausibility (human agreement), faithfulness and runtime complexity.

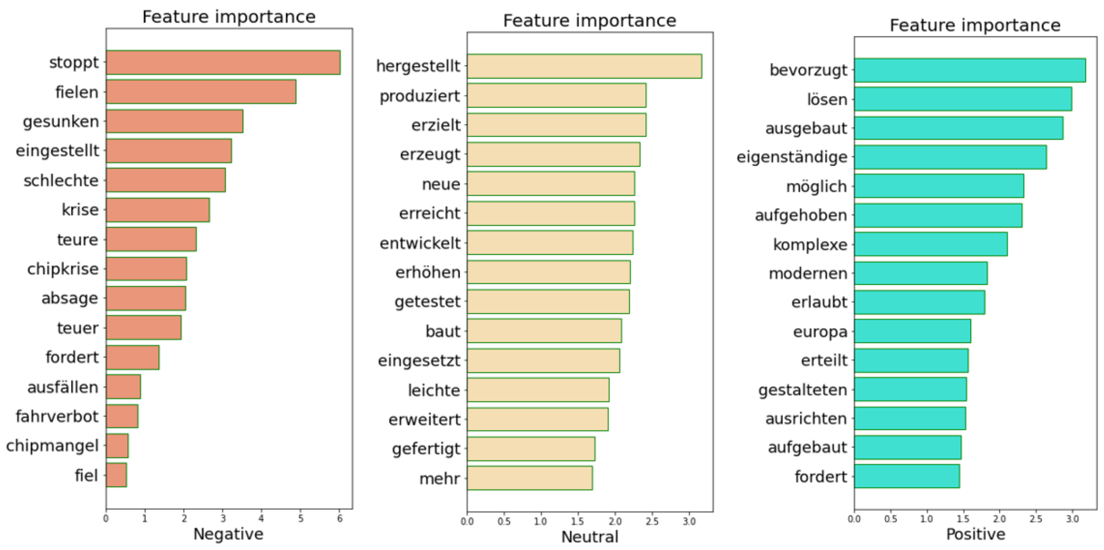


Figure 5: Global Feature Importance Score per Class generated by SHAP and the BERT Sentiment Classifier.

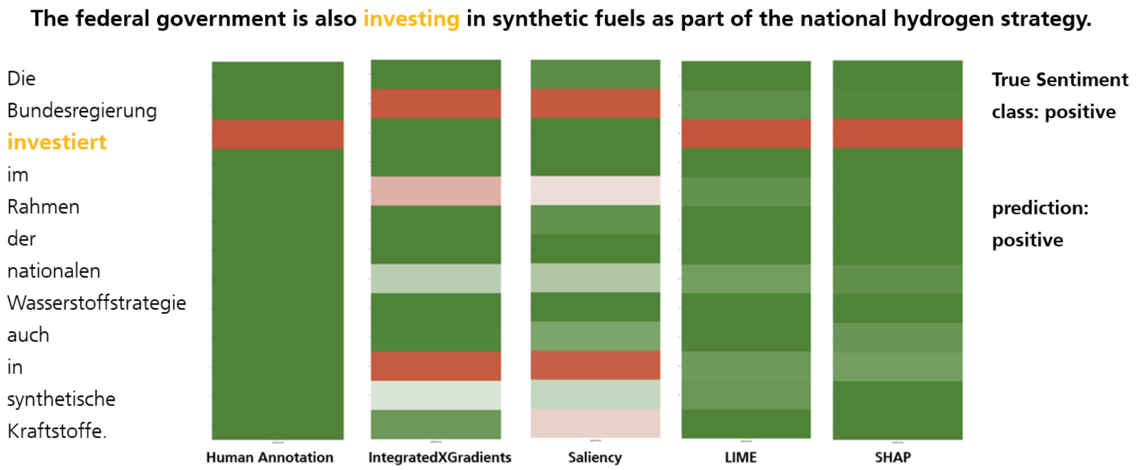


Figure 6: Explainability information generated by different XAI methods for the BERT Sentiment analysis model.

5.1 Results for Human Agreement Performance

Table 6 shows the Human Agreement performance of each explanation algorithm on our test dataset. Regarding plausibility, there is only modest agreement of explanations generated by XAI methods versus human rationales, independent of the ML model. Since rationales are generally very short, the model has a high chance of missing the main evidence in the text. Simpler models like CNN performed slightly better, so we observe a negative correlation between model performance and agreement with human rationales.

XAI methods usually produce a list of top-k features, ordered according to their saliency. However, hyperparameter setting and choice of the optimal cutoff threshold of the ranked feature list impacts the output results. To this aim, we calculated plausibility with a moving threshold t in the range [0.01, 0.2], varying the number of tokens being extracted as salient features, and selecting thresholds that result in the highest F1 score. Related work generally determines the threshold value heuristically based on the length ratio of the gold annotations or based on the saliency weights, as suggested by Yu et al. (2019). Interestingly, the results differ from the heuristically determined threshold based on the length ratio (see Table 2), and suggest that better results could be obtained by choosing a higher threshold, i.e, *negative* mean: 11.7%; std: 4.7, *neutral* mean: 12.3%; std: 5.5, *positive* mean: 11.4%; std: 5.7.

Model	Token F1	IOU
CNN [LIME]	0.066	0.036
LSTM [LIME]	0.062	0.034
BERT [LIME]	0.069	0.037
CNN [IG]	0.044	0.023
LSTM [IG]	0.068	0.037
BERT [IG]	0.039	0.021
CNN [Saliency]	0.062	0.033
LSTM [Saliency]	0.068	0.037
BERT [Saliency]	0.044	0.023
CNN [SHAP]	0.074	0.041
LSTM [SHAP]	0.071	0.039
BERT [SHAP]	**0.078**	**0.043**

Table 6: Evaluation Plausibility of the explainability techniques measured with Intersection over Union (IOU) and Token F1 Score with the gold human annotations.

5.2 Results for Faithfulness

In our use case, we require faithful explanations that are actually relevant to the model's prediction and inner workings. The sufficiency of rationales, based on the AUC score is shown in Table 7. A faithful rationale should display low sufficiency values, as the prediction should be highly influenced by the tokens selected as most salient. As can be seen, simpler model architectures achieve the highest faithfulness scores in terms of sufficiency, with CNN performing best.

	CNN	LSTM	BERT
LIME	**27.29**	37.84	52.66
IG	29.92	37.44	46.15
Saliency	29.32	36.85	46.71
SHAP	27.80	32.66	46.46

Table 7: Comparison of the Faithfulness of the explainability methods measured by AUC for thresholds $\in [0, 10, 20, .., 100]$ defined as the average difference of the AUC before and after masking the top k% words with zero padding. Lower scores are better.

5.3 Computational Efficiency

We also compare the computational efficiency of our XAI techniques. The wait time for an explanation should not be a bottleneck for the task workflow. We recorded the computational time to generate salient features (including visualization heatmaps) for our test dataset. As shown in Table 8, the computational time for generating explanations is particularly high for SHAP in combination with LSTM or BERT.

Explainability Model	BERT	LSTM	CNN
LIME	1549	4744	308
IG	498	40	63
Saliency	486	42	59
SHAP	13362	53917	1245

Table 8: Computational efficiency. We report the mean speed in seconds for generating a saliency map for the test dataset

6 Final Discussion

In this paper, we introduced a new benchmark corpus compiled from online news articles, annotated by 3 domain experts for document-level sentiment. Moreover, it contains multiple rationales that provide evidence for the annotators choice of the overall sentiment. Since many articles have a mixed sentiment, including borderline cases that are difficult to classify, highlighting positive as well as negative aspects mentioned in one single article, yields increased transparency.

We used the corpus as a benchmark for the German language, where resources for studying explainability are scarce. We investigated several ML architectures for the task, in combination with different post-hoc explainability methods. Since there is no single solution that is best suited to every use case, our analysis allows identifying the strengths and limitations of each method.

Our findings indicate that BERT yields the best performance in terms of sentiment classification accuracy. In combination with SHAP, it offers a global view of feature importance, which helps detecting spurious features and bias. We think that end users will profit from XAI methods which allow to get an aggregated view of feature importance for a particular topic category, or based on a specific time frame. However, due to the high dimension of our data, local explanations are overall not very plausible, regardless of the underlying ML model and explainability method. Moreover, the BERT model is less faithful than CNN and LSTM, due to high complexity of the model. For our use case, the computational time for generating explanations with LIME, IG or Saliency would be acceptable in a real-time application, except for SHAP which suffers from a long computational time.

In future work, we seek to identify the training data points responsible for model misclassifications and find training instances that show bias through influence functions (Koh and Liang, 2017), and investigate the impact of the pretrained embeddings and model.

7 Acknowledgements

We thank our students Ina Weidinger and Vincent Hahne for supporting us in the annotation task for the InfoBarometer dataset as well as for helping us to set up the annotation guidelines.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Julia El Zini, Mohamad Mansour, Basel Mousi, and Mariette Awad. 2022. On the evaluation of the plausibility and faithfulness of sentiment analysis explanations. In *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part II*, pages 338–349. Springer.

Shilpa Gite, Hrituja Khatavkar, Ketan Kotecha, Shilpi Srivastava, Priyam Maheshwari, and Neerav Pandey. 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, 7.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Eleni Kalamara, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia. 2022. Making text count: economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37(5):896–919.

Daisuke Katayama, Yasunobu Kino, and Kazuhiko Tsuda. 2019. A method of sentiment polarity identification in financial news using deep learning. *Procedia Computer Science*, 159:1287–1294.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Abby Levenberg, Stephen Pulman, Karo Moilanen, Edwin Simpson, and Stephen Roberts. 2014. Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*, 3(2):109–115.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Anna Marszal et al. 2022. What news can really tell us? evidence from a news-based sentiment index for financial markets analysis. Technical report, Narodowy Bank Polski, Economic Research Department.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8:131662–131682.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Nadia Nedjah, Igor Santos, and Luiza de Macedo Mourelle. 2022. Sentiment analysis using convolutional neural network via word embeddings. *Evolutionary Intelligence*, 15(4):2295–2319.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Lina Maria Rojas-Barahona. 2016. Deep learning for sentiment analysis. *Language and Linguistics Compass*, 10(12):701–719.

Jürgen Schmidhuber, Sepp Hochreiter, et al. 1997. Long short-term memory. *Neural Computing*, 9(8):1735–1780.

Kazuhiro Seki, Yusuke Ikuta, and Yoichi Matsubayashi. 2022. News-based business sentiment and its properties as an economic index. *Information Processing & Management*, 59(2):102795.

L Shapley. 1953. Quota solutions op n-person games1. *Edited by Emil Artin and Marston Morse*, page 343.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *ArXiv*, abs/2010.00389.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. 2021. Dutrust: A sentiment analysis dataset for trustworthiness evaluation. *arXiv preprint arXiv:2108.13140*.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. A fine-grained interpretability evaluation benchmark for neural nlp. *arXiv preprint arXiv:2205.11097*.

Daniel Weiss and Fabian Nemeczek. 2022. A media-based innovation indicator: Examining declining technological innovation systems. *Environmental Innovation and Societal Transitions*, 43:289–319.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *ArXiv*, abs/2102.12060.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Mo Yu, Shiyu Chang, Yang Zhang, and T. Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Conference on Empirical Methods in Natural Language Processing*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

A Appendix

A.1 Query Synonyms

q1: ["Elektromobilität","Elektroauto","Stromer", "Elektrofahrzeug", "BEV","PHEV", "Electromobility", "e-mobility", "e-car", "electric car","emobility", "e-mobilität", "Emobilität"]

q2: ["autonomes Fahren","autonomes Auto","autonomes Fahrzeug","selbstfahrendes Auto", "selbstfahrendes Fahrzeug","fahrerloses Fahren","fahrerloses Auto","fahrerloses Fahrzeug", "autonomous driving","Robotaxi", "autonomes Taxi", "autonomes Shuttle", "autonome Mobilitätsdienstleistungen","robotaxi"]

q3: ["Wasserstoff", "Hydrogen", "FCEV", "Brennstoffzelle", "Wasserstofffahrzeug", "Wasserstoffauto", "Brennstoffzellenfahrzeug", "Brennstoffzellenauto", "fuel cell", "hydrogen", "Power-to-X"]

q4: ["Diesel"]

q5: ["SynFuels", "SynFuel", "Synthetische Kraftstoffe", "Syn-Fuels", "EFuels", "E-Fuels", "alternative Kraftstoffe"]

A.2 Implementation Details

A.2.1 SVM

We utilize a linear svm model operating on bag of word features weigthed by tf.idf, and lowercasing all words. To optimize the models, we employ full-batch gradient descent with L1 regularization on the weight matrices.

A.2.2 CNN

For CNN, we employ an embedding dimension of 300, which is initialized by the GloVe embeddings. The batch size is 64, with a 0.1 dropout rate. We utilize the AdamW optimizer with a learning rate of 5*e*-5. Window sizes $\in [2, 3, 4, 5]$ were selected, with 100 out channels. Regarding stride, we keep the hyperparameter at the default value 1. CNN models are trained using an early stop over the validation set and up to 50 training epochs.

A.2.3 LSTM

For LSTM, we employ an embedding dimension of 300, which is initialized by the GloVe embeddings. The batch size is 64, and the dropout rate 0.1. We use a learning rate of 5*e*-5 along with the AdamW optimizer. 4 LSTM layers were selected, with a hidden layer size of 300. LSTM models are trained using an early stop over the validation set and up to 50 training epochs.

A.2.4 Transformer

As a base for all our experiments we use the German BERT-BASE model which consists of 12 layers, a hidden state size of 768 dimensions per token amounting to a total of 110 million parameters. The parameters of this model are initialized using bert-base-german-cased, which has been released by deepset.ai. We trained the model with a learning rate of 5*e*-5. We chose the best model using early stopping with the best number of epochs determined by using the validation splits.