HWWI Working Paper Series

Inefficient forecast narratives: A BERT-based approach

Alexander Foltas

HWWI Hamburgisches WeltWirtschaftsInstitut

HWWI Working Paper No. 6/2025

Authors:

Alexander Foltas (alex.foltas@gmx.de)
Helmut-Schmidt-Universität Hamburg

Imprint

Publication Series: HWWI Working Paper Series, ISSN 2750-6355
Responsible editor: Michael Berlemann

Hamburg Institute of International Economics (HWWI)
Scientific Director: Prof. Dr. Michael Berlemann
Managing Director: Dr. Dirck Süß
Mönkedamm 9 | 20457 Hamburg | Germany
Phone: +49 40 340576-0 | Fax: +49 40 340576-150
info@hwwi.org | www.hwwi.org

Inefficient forecast narratives: A BERT-based approach

Alexander Foltas[1]

[1] *Helmut-Schmidt-University Hamburg, Holstenhofweg 85, 22043 Hamburg, Germany.*

Abstract

This paper contributes to previous research on the efficient integration of forecasters' narratives into business cycle forecasts. Using a Bidirectional Encoder Representations from Transformers (BERT) model, I quantify 19,300 paragraphs from German business cycle reports (1998-2021) and use them to predict the direction of consumption forecast errors. By testing the model on an evaluation sample, I find a highly significant correlation of modest strength between predicted and actual sign of the forecast error.
The correlation coefficient is substantially higher for 12.8% of paragraphs with a predicted class probability of 85% or higher. By qualitatively reviewing 150 of such high-probability paragraphs, I find recurring narratives correlated with consumption forecast errors. Underestimations of consumption growth often mention rising employment, increasing wages and transfer payments, low inflation, decreasing taxes, crisis-related fiscal support, and reduced relevance of marginal employment. Conversely, overestimated consumption forecasts present opposing narratives. Forecasters appear to particularly underestimate these factors when they disproportionately affect low-income households.

Keywords: Macroeconomic forecasting; Evaluating forecasts; Business cycles; Consumption forecasting; Natural language processing; Language Modeling; Machine learning; Judgmental forecasting

1 Introduction

In his widely acclaimed work, "Narrative Economics," Shiller (2017) argues that "popular sense-making stories" play a crucial role in explaining human decision-making and demonstrates their substantial utility in explaining economic phenomena such as recessions or bubbles in financial markets. Since then, a growing number of economists have emphasized the role of narratives in economic activity. In their recent survey of narrative economics, Roos and Reccius (2024) enhance conceptual clarity within the field by proposing the term "collective economic narratives," which they define as "a sense-making story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions."

Narratives are also crucial for professional macroeconomic forecasting. While forecasters base their numeric point forecasts on econometric or statistical models, they usually adjust their model results to consider information not captured by traditional indicators, such as announced policy measures or data scarcity (Fildes and Stekler 2002). Although forecast adjustments based on human judgment are subjective, they are not arbitrary. Instead, they are reasoned with economic narratives, which can be found in business cycle reports accompanying numerical point forecasts.

Despite the potential for bias and inefficiency in judgmental forecast adjustments and their underlying narratives, they are considerably less investigated than models relying solely on quantitative data. Stekler and Symington (2016) conduct one of the first in-depth analyses of economic narratives, which they call "qualitative forecasts," in a case study of the FOMC minutes around the Great Recession. Remarkably, the authors find that the FOMC committee correctly assessed the implications of the housing market decline and the volatility of the financial markets and their associated risks throughout 2007, yet they failed to forecast the recession and were even late in recognizing it. While Stekler and Symington (2016) provide several possible explanations for why

the forecasters failed to correctly incorporate their qualitative information set into forecasts, their study highlights the necessity for systematic evaluations of forecasters' economic narratives and their incorporation into numeric point forecasts.

Recent substantial methodological advances in natural language processing (NLP) provide researchers with increasingly powerful tools for quantifying economic narratives, allowing for a systematic analysis of their informational value. Clements and Reade (2020) conduct a sentiment analysis of the Bank of England's Quarterly Inflation Reports and find that the reports' tonality can predict forecast errors, suggesting that forecasters inefficiently utilize their qualitative information set for growth and inflation forecasts. Müller (2022) and Sharpe, Sinha, and Hollrah (2023) confirm that tonality is inefficiently incorporated into growth and inflation predictions for German professional forecasters and the Federal Reserve. Furthermore, Foltas and Pierdzioch (2022b) use topic modeling to show that business cycle reports' thematic compositions can predict growth and inflation forecast errors and suggests that significant topics contain inefficiently utilized qualitative information (see also Foltas 2022; Foltas and Pierdzioch 2022a).

Previous quantification-based approaches represent important contributions, indicating that forecasters systematically struggle to exploit the informational value of their reports. However, these studies fail to provide full numeric representations of economic narratives. Theme and tonality are aspects of sense-making stories instead of full narratives themselves. Most of the cited studies focus on one of these aspects while neglecting the other, resulting in inflation topics that do not allow for distinctions between positive or negative sentiment, or sentiment indices that cannot differentiate between inflation or unemployment. Furthermore, researchers do not quantify forecasters' reasoning for their judgments, thus ignoring the "sense" of the stories. Forecasters might have vastly different rationales for the same sentiment toward enacted stimulus packages or arranged trade agreements, with different associated forecast errors.

Therefore, I extend previous research using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin *et al.* 2018) and continue its training on business cycle reports to obtain context-aware word representations. BERT's word representations fully quantify all characteristics of its underlying textual data, differentiating them from measurements employed in previous research, such as sentiment or topics. BERT represents a milestone for NLP techniques due to its far superior performance compared to earlier approaches (Zhao *et al.* 2023).

Using BERT's word representations, I test the textual forecast efficiency (Foltas and Pierdzioch 2022a) of German consumption forecasts (1998-2021). Consumption forecast efficiency is relatively under-researched despite consumption being the largest subcomponent of GDP. Furthermore, Eicher *et al.* (2019) find that private consumption is by far the most important error contributor to IMF growth forecasts in crisis times using a sample of 84 countries. Tsuchiya (2021) confirms that private consumption forecast errors have the highest contribution to growth forecast errors in crisis times using World Bank forecasts. Tsuchiya (2021) also finds an even stronger impact of consumption forecast errors during non-crisis periods, concluding that improvements in private consumption forecasts are required to reduce forecast errors and the upward bias of growth forecasts.

Previous studies investigating consumption efficiency mostly rely on the widely used Mincer-Zarnowitz regression (Mincer and Zarnowitz 1969), which utilizes a quadratic loss function, and do not examine forecasters' narratives. Using this approach, Dovern and Weisser (2011) find efficient private consumption forecasts for Germany and other G7 countries, although they confirm an upward bias for German consumption forecasts. Similarly, Eicher *et al.* (2019) and Tsuchiya (2021) find efficient IMF and World Bank private consumption forecasts. To the best of my knowledge, only Deschamps and Bianchi (2012) reject the efficiency of Chinese consumption forecasts. However, multiple studies question the validity of Chinese-reported economic data (e.g. Keidel 2001, Wu 2007, Clark, Pinkovskiy, and Sala-i-Martin 2020), limiting the conclusions that can be drawn.

The Mincer-Zarnowitz regression is widely popular, despite studies finding asymmetric loss

functions (e.g. Krüger and Hoss 2012, Pierdzioch, Rülke, and Tillmann 2016). As the loss function is still subject to debate, I follow the approach of Patton and Timmermann (2007) to test forecast efficiency under flexible or unknown loss functions. In their Proposition 3, Patton and Timmermann (2007) show that forecast efficiency can be rejected if a set of predictor variables X_t at period t has significant explanatory power for the sign of the forecast error. Hence, the forecast efficiency can be tested by estimating a classification model of the form: $1_{e_{t+1>0}} = \alpha + \beta X_t + u_{t+1}$, where $1_{e_{t+1>0}}$ is the indicator function of the forecast error, β denotes a vector of coefficients, and u_{t+1} is an error term. Forecasts are considered optimal when the null hypothesis $\beta = 0$ cannot be rejected.

Most previous studies of economic narratives in forecasting provide little insight into how forecasters might improve their forecasts, a result of heavy abstractions, the usage of black-box methods, and imperfect quantification methods. I aim to overcome this problem by training the BERT classifier on text chunks that roughly include a single paragraph of a forecast report instead of the full document. This method lowers the classifier's overall performance since a high number of paragraphs likely contain no textual information with value for consumption forecast errors. However, chunking provides class probabilities for each individual paragraph, and particularly, paragraphs with high probabilities might provide insights into underutilized textual information. Therefore, I will investigate and compare low-loss chunks to specify underutilized economic narratives and support my findings with textual snippets from the corpus.

The remainder of this paper is organized as follows: In Section 2, I briefly explain text classification with language models, present the BERT model architecture, and the details of my fine-tuning approach. Section 3 presents the corpus and data transformations. In Section 4, I analyze the model's performance and compare it to benchmark models and various subsamples. I then investigate and present economic narratives in Section 5 and conclude in Section 6.

2 Methods

2.1 Text Classification

Text classification is a standard task of NLP, aiming to assign classes or labels to textual units such as sentences, queries, paragraphs, or documents. Among the most common applications of text classification are sentiment analysis, news categorization, topic analysis and spam detection (Minaee *et al.* 2022).

The most successful text classification approaches utilize deep learning-based language models (LMs), which model the generative likelihood of word sequences (Minaee *et al.* 2022). A central pillar in the development of most modern LMs is the Transformer model architecture (Vaswani *et al.* 2017), which converts word sequences into contextualized word representation vectors. These vectors capture semantic features using six layers of neural networks (encoders) and reconvert the word vectors through six layers of decoders, with applications including translation and question answering. Encoders and decoders are complex architectures consisting of multiple sublayers that capture the relationships and meanings of words within a given context (see Vaswani *et al.* (2017) for a detailed explanation). At the time of its development, the Transformer surpassed all state-of-the-art translation models for a fraction of the training costs, thus becoming the foundational model for a range of pre-trained language models (PLMs) (Zhao *et al.* 2023).

One of the most successful Transformer-based PLMs is the BERT architecture (Devlin *et al.* 2018). BERT operates by transforming textual inputs into tokens using the SentencePiece tokenizer (Kudo and Richardson 2018). The tokenizer uses a predefined dictionary consisting of the most frequently used words and subwords, splitting less familiar words accordingly. For example, a BERT tokenizer might split the textual input "growing economy" into the tokens "grow," "##ing," and "economy." The model links each token to a 768-dimensional word vector, which is contextualized using 12 layers of encoders.

BERT's word vectors and encoder parameters are obtained through two steps of unsupervised pre-training. First, the model predicts a percentage of masked tokens in its corpus based on the surrounding tokens (masked language modeling), continuously updating its parameters to minimize loss with a portion of the corpus functioning as a test sample to prevent overfitting. Second, BERT is presented with two sentences and guesses whether the second sentence follows the first one (next sentence prediction). In both steps, BERT learns to "understand" the relationships between tokens and quantifies them into contextualized word representations.

The word representations of a pre-trained BERT model can be fine-tuned to solve a variety of NLP downstream tasks (see Aftan and Shah 2023 for a survey of BERT applications). Fine-tuning for text classification typically involves attaching a simple linear neural network (LNN), which sorts textual inputs into classes according to their pooled word representations. The model is then trained in a supervised manner on a training corpus, updating both the main model's and the LNN's parameters to minimize the classification loss. While fine-tuning is relatively simple and computationally inexpensive compared to pre-training, it delivers outstanding performance (Gasparetto *et al.* 2022; Minaee *et al.* 2022), leading to high flexibility for a pre-trained BERT model.

The current state-of-the-art LMs are Large Language Models (LLMs), which primarily differ from PLMs by their number of parameters[1] and training data. LLMs are a major breakthrough in AI development due to their remarkable ability to solve complex problems through in-context learning (ICL) abilities. ICL refers to the phenomenon where a trained LLM learns new tasks simply by being provided with a few examples, without the necessity of fine-tuning or any parameter updates. LLMs learn by analogies, which mimic the decision process of human beings. The ICL capabilities of LLMs were unforeseen by most researchers, and their exact working mechanisms remain subjects of debate (Dong *et al.* 2022; Min *et al.* 2022; Zhao *et al.* 2023).

Despite the astonishing capabilities of LLMs, it is doubtful whether ICL is sufficient for the purposes of this study. Typically, researchers conduct ICL on tasks that are immediately solvable for humans. Relating text segments to the direction of a forecast error is not one of these tasks. Furthermore, LLMs are especially data-hungry, making it unfeasible to fine-tune an LLM with the corpus of this study. In Section 4.1, I will compare the classification performance of a fine-tuned BERT model with two LLMs.

2.2 The Approach

This study employs the "bert-base-german-dbmdz-uncased" model[2] developed by the Munich Digitization Center (MDC), which was trained on 16 GB of data from various sources, including Wikipedia, EU publications, movie subtitles, and newspaper articles. Multiple studies have shown that the performance of pre-trained language models (PLMs) can degrade when applied to narrow domains that differ substantially from their training corpus (Thompson *et al.* 2019). Therefore, I enhance the performance of the MDC BERT model through domain adaptation (Guo and Yu 2022). First, I expand the MDC tokenizer's dictionary by adding the 600 most frequently used tokens from my corpus that were not included in the original dictionary of around 30,000 tokens. These newly added tokens include key economic concepts such as "bruttoinlandsprodukt" (gross domestic product), "arbeitslosigkeit" (unemployment), and "konsumausgaben" (consumption expenditure). The original tokenizer splits these terms into multiple tokens (e.g., "arbeitslos" and "##igkeit"), which could impede further analysis.

Then, I continue pre-training the MDC BERT model on my corpus using masked language modeling with the Transformers Python library[3], masking 15% of the tokens, which is the default setting. Using the same library, I fine-tune the model to predict the sign of the forecast error $1_{e_{t+1>0}}$

1. The BERT PLM consists of 330 million while the GPT-3 LLM uses 175 billion parameters(Zhao *et al.* 2023).
2. See https://huggingface.co/dbmdz/bert-base-german-uncased
3. See https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForMaskedLM

with a LNN[4]. The forecast error e_{t+1} is defined as $e_{t+1} = c_{t+1} - \hat{c}_{t+1|t}$, where $\hat{c}_{t+1|t}$ being forecasters' one-year-ahead annual consumption growth forecast at year t and c_{t+1} the actual outcome. Both domain adaptation and fine-tuning follow common practice by using 70% of the chunks for training and 30% for evaluation.

As with most deep-learning approaches, domain adaptation and fine-tuning require a number of previously set hyperparameters, which are crucial for the model's performance. Unfortunately, extensive hyperparameter optimization comes with substantial computational costs (Izsak, Berchansky, and Levy 2021). Therefore, I restrict hyperparameter tuning to the three most crucial parameters: gradient steps (frequency of parameter updates during training), learning rate (size of parameter updates), and weight decay (a penalty to large parameters to avoid overfitting). I tune the hyperparameters automatically using Bayesian Optimization and Hyperband (BOHB) with the Wandb Python library[5] (Falkner, Klein, and Hutter 2018).

To benchmark the fine-tuned BERT's classification performance, I employ OpenAI's[6] LLMs "text-embedding-3-small" and "text-embedding-3-large." The models encode textual inputs into embedding vectors of 1536 (text-embedding-3-small) and (text-embedding-3-large) 3072 dimensions. I employ the models to encode each chunk, use the resulting vectors to train LNN and random forest classifiers (Breiman 2001), and compare their performance to the fine-tuned BERT model. As OpenAI's models are not open source, it is not possible to adapt them to my corpus or tune their hyperparameters.

To evaluate the classifiers' performances on their 30% test sample, I use the Matthews correlation coefficient (MCC). MCC is calculated directly from the confusion matrix as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. MCC ranges between -1 and 1 and is similar to Pearson correlation coefficients in its interpretation. This metric is more reliable than popular alternatives as accuracy or f1-score in binary classification (Chicco and Jurman 2020), as it includes all values of the confusion matrix, preventing misleading or overoptimistic scores on imbalanced data sets.

To evaluate the text classifiers' statistical significance, I test the null hypothesis that the signs of the forecast errors are independent of their textual reports using a permutation test. The permutation test is effective for assessing classification performance when the number of input features far exceeds the number of data points, making it widely applied in data mining and machine learning. The test measures the likelihood that the observed test metric could be obtained by chance by randomly reshuffling the data set's classes without replacement to create multiple additional samples. A classifier is then trained for each permuted sample, requiring fine-tuning of the BERT model each time. Due to the purely random nature of the drawn samples, the newly trained classifiers should have no predictive value for the permuted classes. A p-value represents the fraction of random data sets with classification models that achieve an equal or higher MCC than the original data (Ojala and Garriga 2010).

3 Data

I study 412 consumption forecasts from 1998 to 2021 from the IWH forecasting dashboard (Heinisch *et al.* 2023) for the following year. These forecasts cover the following year and are sourced from publications by three major German research institutes: the German Institute for Economic Re-

4. See https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification
5. See https://docs.wandb.ai/guides/sweeps
6. See https://openai.com/

Table 1. Descriptive Statistics

		Forecast Errors			N Tokens	
Format	N	Median	MAD	FE > 0	Median	MAD
Documents	412	-0.39	0.59	0.28	5884	2822
Chunks	19284	-0.39	0.55	0.27	135	42

Notes: N = Number of documents, chunks, or tokens. MAD = Median absolute deviation. FE > 0 = Proportion of positive forecast errors.

search (DIW), the ifo Institute for Economic Research (IFO), and the Kiel Institute for the World Economy (IFW). Additionally, I include forecasts from smaller institutes: the Hamburg Institute of International Economics (HWI), the employer-related German Economic Institute (IWK), the trade union-related Macroeconomic Policy Institute (IMK), and the Joint Economic Forecast (GED), which is a joint economic analysis by five leading German research institutes.

Table 1 shows the descriptive statistics for the original state of the forecasts, pairing each forecast with its associated business cycle report, and for the corpus' division into chunks (explained below). The forecasters' median private consumption forecast for the following year is 0.39 percentage points above the realized value[7]. Furthermore, nearly three-quarters of the sample's forecasts exhibit a negative sign, indicating that consumption forecasters either have a substantial bias or their loss function is highly asymmetric.

The median document has 5,884 tokens after text cleaning. Text cleaning involves a supervised exclusion of irrelevant parts of the business cycle reports concerning German aggregated private consumption, such as sections on other countries' economies, methodology, and extended micro-economic analysis. Sections indirectly related to aggregated consumption, such as analysis on exports or aggregated investments, are retained to allow for unexpected findings and to avoid further reducing the corpus size. Additionally, I automatically correct common optical character recognition errors, remove footnotes and seldom-used special characters, unify tokens affected by syllabification, and uncase letters. Finally, all dates and other numerals are replaced with "[DATE]" and "[NUM]" tokens to prevent the classifier from training on quantitative information.

Since the number of tokens exceeds BERT's 512 token limit for all documents, I split the documents into text chunks containing one paragraph by searching for the pattern ".\t" or ".[NUM]\t", accounting for German citation rules. Model performance improves when using chunks of similar size. Therefore, I set a minimum chunk size of 64 tokens and a maximum of 256 tokens. Paragraphs exceeding the upper limit are split into 2-3 similarly sized chunks while avoiding breaking sentences. Paragraphs below the lower limit are dismissed since they are unlikely to include substantial information on narratives, resulting in a 4.4% loss of tokens.

The documents vary significantly in length, as indicated by the median absolute deviation of 2,822 tokens. Consequently, splitting the corpus into chunks might overrepresent forecast errors in longer documents since each chunk inherits the document's forecast error. However, as Table 1 shows, the impact on the forecast errors' descriptive statistics is negligible. Table 2 provides a breakdown of the corpus by institute. The median number of tokens per document ranges from 5,116 to 7,847 for most institutes, with 820 tokens (HWI) and 14,040 tokens (GED) as outliers. These business cycle reports process qualitative information either more generally or in greater detail compared to the majority of reports.

The smaller institutes (IWK, IMK, and HWI) forecast private consumption with higher accuracy

7. Due to high negative forecast errors during the Great Recession and the Covid pandemic, the average forecast error is -0.95 with a range from -8.11 to 1.91. I present the median because the forecast errors' range does not impact further analysis, and the low average might be misleading.

Table 2. Descriptive Statistics by Institute

		Forecast Errors			N Tokens	
Institute	N	Median	MAD	FE > 0	Median	MAD
DIW	68	-0.47	0.67	0.26	5116	1530
GED	48	-0.38	0.74	0.23	14040	1526
HWI	60	-0.28	0.50	0.33	820	98
IFO	47	-0.43	0.63	0.26	7847	1871
IFW	86	-0.67	0.58	0.14	7153	2075
IMK	65	-0.25	0.49	0.42	4568	1773
IWK	38	-0.17	0.62	0.45	6571	1056

Notes: N = Number of documents, chunks, or tokens. MAD = Median absolute deviation. FE < 0 = Proportion of positive forecast errors.

than the rest of the sample. This is notable given that the analysis and policy recommendations of the IMK and the IWK often contradict each other. However, the lower absolute forecast error difference might be due to differences in their business cycle publication months, as the smaller institutes seldom publish in the first quarter[8].

4 Model performance

4.1 Domain adaptation and fine-tuning

Figure 1 presents the loss during continued training for domain adaptation with an extended vocabulary on both the training and evaluation samples. Initially, the model achieves lower loss on the evaluation sample than on the training sample. During training, each parameter has a 10% dropout probability, which sets the parameter to zero to reduce overfitting (El Anigri, Himmi, and Mahmoudi 2021). This means an already trained model should typically show lower loss on its evaluation sample. During the first 1,000 steps (parameter updates), the model's masked language modeling (MLM) loss improves smoothly. After this point, the evaluation loss curve flattens, leading to the training loss sinking below the evaluation loss, which can be considered a sign of overfitting. I select the last epoch prior to the intersection of the loss curves, which occurs after nine epochs (full iterations of the training data) with 1,117 steps, a training loss of 1.06, and an evaluation loss of 1.03.

Fine-tuning the adapted BERT model for text classification shows overfitting by the second epoch, which is expected for a smaller corpus. Therefore, I train the model for only one epoch (211 steps), resulting in a training loss of 0.57 and an evaluation loss of 0.53.

Table 3 presents BERT's classification performance on its evaluation sample and compares it to OpenAI's benchmark models. With an MCC of 0.234, the model demonstrates a moderate correlation between the predicted and actual direction of forecast errors. The model's accuracy falls below that of a naive model that always predicts negative forecast errors, which would score around 0.73. Approximately 45% of the model's positive forecast error predictions are correct, and it retrieves the correct class at a similar rate.

It is not possible to fine-tune the embedding model's of OpenAI. Instead, I utilize their latest embedding LLMs, namely "text-embedding-3-small" and "text-embedding-3-large," to quantify each text chunk and then train a classification model on the output vectors. Initially, I attempt to classify using the classification head provided by the transformers python library, which is also

8. See Döhrn and Schmidt 2011 for the impact of forecast horizons on forecasting accuracy.

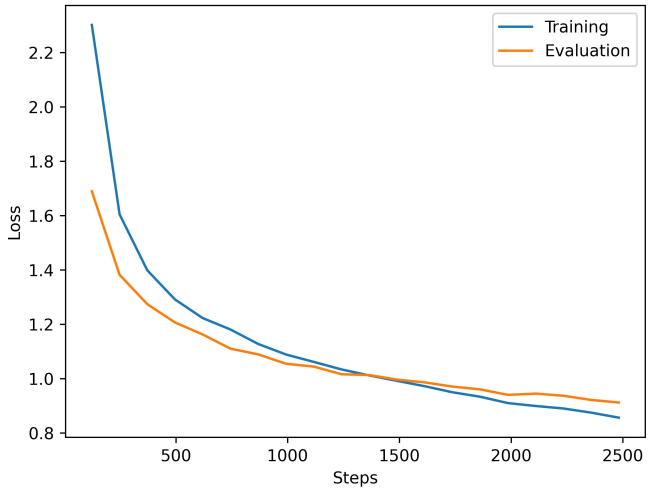


Figure 1. Domain adaptation loss

Table 3. Model performance

Classifier	Model	MCC	Accuracy	Precision	Recall	P-value
Fine-tuning	BERT	0.234	0.684	0.452	0.460	0.000
LNN	BERT	0.000	0.729	0.000	0.000	-
	text-embedding-3-small	0.000	0.729	0.000	0.000	-
	text-embedding-3-large	0.000	0.729	0.000	0.000	-
Random forest	BERT	0.075	0.731	0.694	0.159	0.000
	text-embedding-3-small	0.061	0.730	0.846	0.007	0.000
	text-embedding-3-large	0.069	0.731	0.750	0.011	0.000

Notes: LNN = Linear neural network; MCC = Mathews Correlation Coefficient

used in the fine-tuned BERT model. This classification head consists of a LLN for predictions and includes a dropout layer to mitigate overfitting during training.

The performance of the LLN for both benchmark models is inadequate, with no true positives in the evaluation sample. Although the classifier structure for the benchmark models is identical to that of the fine-tuned BERT model, their training methods differ. Fine-tuning updates all parameters within the model, thereby altering the embedding vectors themselves. Conversely, OpenAI's embedding vectors are fixed, so only the LLN parameters are updated. I observe similar performance results when training the LLN on the fixed embedding vectors of the BERT model with domain adaptation. Consequently, using LLNs as classifiers for fixed embedding vectors is inadequate, complicating the comparison between the fine-tuned BERT and state-of-the-art LLMs.

Following OpenAI's text classification guide[9], I proceed to train classification random forests on the embedding vectors of all three models. All models yield comparable MCCs ranging from 0.061 to 0.075, which is notably lower than the MCC achieved by the fine-tuned model. The high precision and low recall suggest that random forest classification predicts significantly fewer positive forecast errors, thus explaining the slightly increased accuracy. Although all three models perform similarly

9. See https://cookbook.openai.com/examples/classification_using_embeddings .

weakly, the domain-adapted BERT slightly outperforms its competitors due to its higher overall performance and improved ability to detect positive forecast errors. Given the inability to achieve comparable classification performance without fine-tuning, it is evident that smaller models still have utility for specific tasks with small datasets.

Despite their low or moderate performances, all models exhibit high significance with p-values of zero, allowing the rejection of the hypothesis of textual forecast efficiency under flexible loss (Foltas and Pierdzioch 2022a). A p-value of zero implies that no model trained on permuted data could outperform the MCC of the true model. Due to the considerable computation times, particularly for fine-tuning, only 100 permuted datasets per model are utilized. However, most permuted models yield an MCC close to zero, indicating no correlation between model prediction and true label in the evaluation set. Among the 400 permuted models, only two exceed an MCC of 0.02, with the highest outlier reaching 0.04.

Given that business cycle reports typically cover a broad range of economic information beyond private consumption, the relatively poor performance of the main model over the full corpus is expected. However, its high significance suggests that text-based classification of forecast error signs is feasible in specific cases, potentially due to inefficiently exploited economic narratives. Consequently, Section 4.2 delves into the performances of various subsamples.

4.2 Subsample performance

Suppose that the corpus C comprises chunks with inefficiently exploited textual information C_N and chunks containing no relevant information for the direction of consumption forecast errors C_U. Then the performance of the overall model MCC_C depends on the performance of relevant chunks MCC_N and the ratio of $p = C_N/C$. Due to $MCC_U = 0$, a thematical subsample of the corpus C_S exhibiting higher performance ($MCC_S > MCC_C$) is expected to contain a relatively greater proportion of chunks with inefficiently utilized textual information ($p_S > p$).

Therefore, I partition the corpus to identify economic domains with a higher frequency of relevant narratives. To achieve this, I utilize a compilation of consumption-related and general economic terms, creating a subset for each of these terms. A subset based on a key term encompasses all chunks containing the term itself or any of its declinations. Given the compound nature of the German language, I also incorporate chunks where the key term appears as a subword. Different subsamples may overlap, with certain instances where one key term subsample entirely encompasses another. For instance, the term "preis" (price) is encompassed both within "ölpreis" (oil price) and "importpreis" (import price), leading to all chunks within the latter two subsets being included in the price sample.

Figure 2 illustrates the key term subsamples that exhibit higher classification performance compared to the overall corpus. Some significant economic terms failed to surpass the MCC threshold of 0.3 or were omitted due to the presence of similar tokens in the figure. The terms have been automatically translated using DeepL[10].

Several key term subsamples demonstrate significantly higher classification performance compared to the overall corpus, highlighting three thematic groups with notable performance. The first group revolves around employment and labor market-related tokens, such as "jobs," "working hours," and "unemployment," as well as terms related to labor compensation like "labor costs" and "wage increase." Given that net disposable income strongly influences private consumption, the prevalence of inefficient information related to this group is plausible.

More specific terms within this group exhibit higher MCC scores, indicating a higher frequency of inefficiently applied textual information within those subsamples. For instance, the "wage increase"

10. The term "employment" appears twice. The subset with fewer chunks is translated as "erwerbstätigkeit" and the other as "beschäftigung"

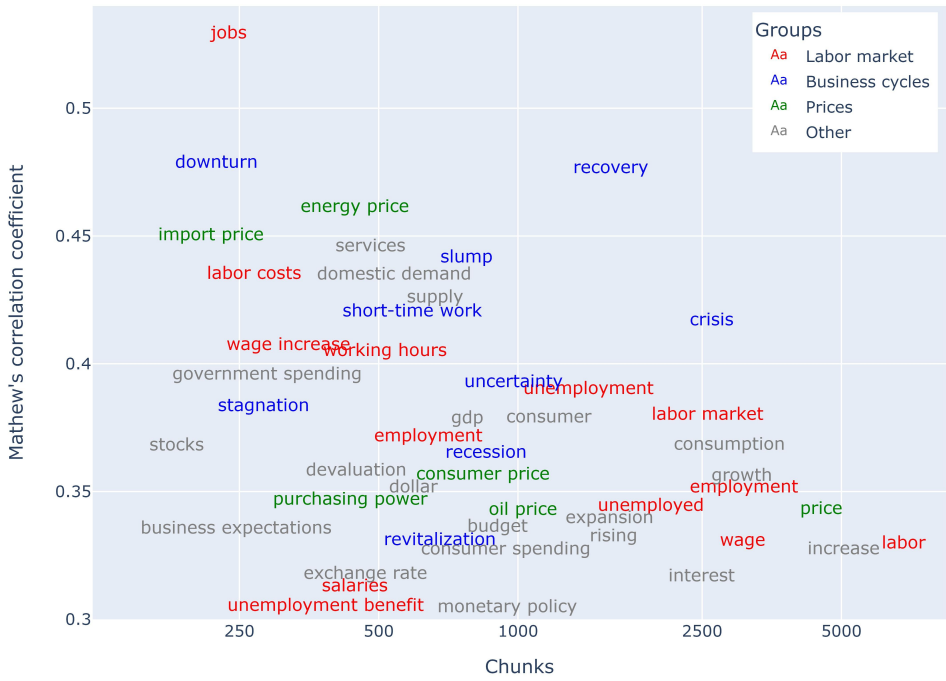


Figure 2. Key term subsample performances

sample demonstrates superior performance to "wage," suggesting that discussions of actual or potential wage raises are particularly prone to inefficiencies.

The second group pertains to crises and business cycle turning points, featuring terms like "downturn," "slump," "recovery," and "crisis." The "recovery" term stands out as it combines a relatively large subsample with high performance, suggesting that the model finds inefficient narratives particularly well during actual or expected economic upswings. This result is plausible as forecasters struggle to forecast turning points (Fildes and Stekler 2002).

The last group focuses on prices, including terms such as "energy price," "import price," and "price." Particularly, the close to 500 chunks containing "energy price" outperform other price-related subsamples. Although oil prices are one of the biggest contributors to energy prices, its subsample performs considerably worse. This discrepancy could be attributed to chunks discussing historical events like the oil crises of the 1970s, which may not be directly relevant to current discussions on energy prices.

While terms like "consumer," "consumption," and "consumer spending" outperform the overall corpus, they fall short of the performance achieved by subsamples focusing on determinants of consumption. Similarly, terms like "services" or "government spending" often appear within the context of other thematic groups, as will be discussed in Section 5.

Notably, "stocks" is the only term related to consumer wealth, a major determinant of private consumption. This suggests either greater success in utilizing wealth-related narratives for consumption forecasting or the model's lesser ability to identify inefficiencies in this area.

After providing a thematic overview of inefficiently exploited information, the following section will examine specific underutilized narratives.

5 Narratives analysis

5.1 Overview

The model's performance is relatively modest when reviewing the whole corpus. However, BERT classifies 12.8% of the sample with a probability of 85% or higher. With an MCC of 0.765, the model's performance on high probability classifications far exceeds the overall performance. While linking a language model's predictions to specific textual patterns is notoriously difficult, investigating high probability classifications might reveal textual patterns related to the sign of a forecast error.

Therefore, I create a subsample of 150 chunks with low loss, which I review to determine their contained economic narratives. To create the subsample, I divide the sample into its classes (sign of the forecast error) and select the 75 chunks with the lowest loss. I skip chunks if the same combination of forecast error and publication year already appears 10 times within the subsample. This approach ensures a diverse composition of various publication years within the subsample, allowing for the highlighting of structural similarities between narratives from different periods.

To provide an overview, I categorize each narrative under a superordinate economic concept such as private consumption, employment, or prices. Figure 3 illustrates the number of chunks with narratives related to the subsample's most frequent concepts, subdivided by their forecast errors. It's important to note that most chunks are counted multiple times as they contain narratives relating to multiple areas. For example, chunks with narratives concerning employment often also contain statements on wages.

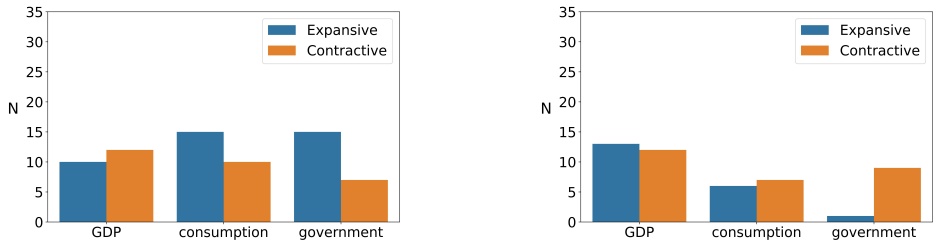
The bar colors subdivide the narratives into "expansive" and "contractive," where "expansive" signifies forecasters observing, discussing, or predicting an increase in measurable variables related to a specific concept, and "contractive" indicates a decrease. The exact criteria for classification vary depending on the concept.

I classify narratives concerning the minimum wage, the euro exchange rate, and tax rates based on literal increases or decreases. For instance, the expansive euro narrative bar counts chunks discussing recent or expected euro revaluations. Government narratives refer to fiscal deficit, thus a contractive government narrative either refers to chunks discussing contractive fiscal policy or shrinking budget deficits.

Macroeconomic variables such as GDP, consumption, or prices are classified as expansive narratives when forecasters consider growth rates as relatively high or anticipate accelerating growth. For instance, a statement like "private consumer spending is likely to increase even more strongly next year" (IMK 2014/12) would be categorized as expansive. Conversely, GDP narratives are labeled as contractive when forecasters perceive growth rates as low or decelerating. For example, "economic momentum slowed noticeably in the first half of the current year. The rate of expansion of gross domestic product halved in the first quarter to [NUM] figure [NUM]" (IFW 2018/06).

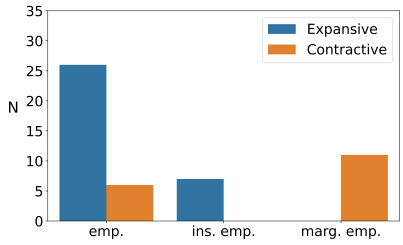
In cases where narratives are unclear or mixed, I classify them based on forecasters' sentiment. For instance, a statement like "private consumption will gradually pick up again" (IFO 2001/12) implies low private consumption but with a rather positive sentiment, thus categorized as an expansion consumption narrative. For GDP, consumption, employment, and insurable employment, expansion and contraction narratives are associated with positive and negative sentiment, while the relationship is reversed for price narratives. Tax narratives do not directly translate into a sentiment category as forecasters' sentiment heavily depends on the institute, economic situation, or specific tax. Similarly, government, marginal employment, wage, minimum wage, transfer income, and euro narratives do not correspond to specific sentiments.

Whether forecasters perceive growth rates as high or low heavily depends on specific economic circumstances. For instance, in the statement "the German economy slipped into a crisis of historic proportions in the winter half-year [DATE] [NUM]. [...] by contrast, private consumer spending remained virtually stable, not least thanks to the stabilizing effect of short-time work and the scrappage scheme as well as the low price increase." (IMK 2009/07), the relative stability of private

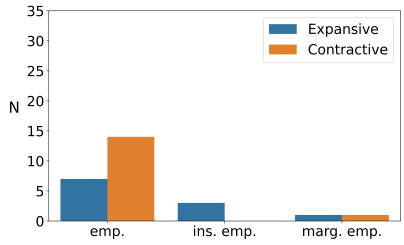


(a) Economic outlook - positive FE

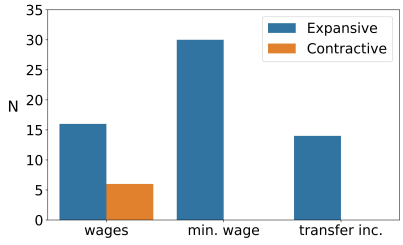
(b) Economic outlook - negative FE



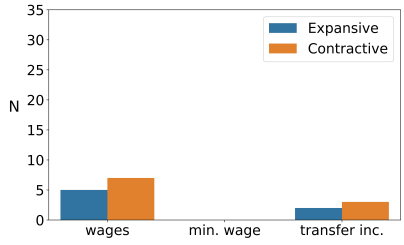
(c) Employment - positive FE



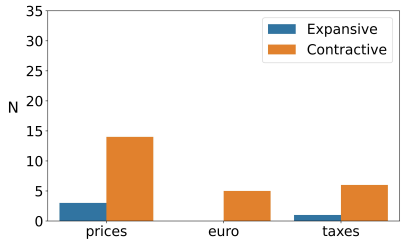
(d) Employment - negative FE



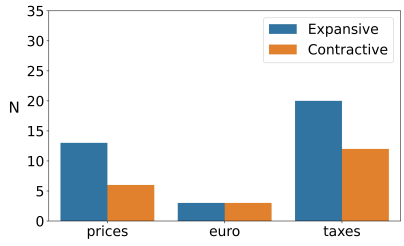
(e) Income - positive FE



(f) Income - negative FE



(g) Price factors - positive FE



(h) Price factors - negative FE

Figure 3. Narratives in the low loss subsample.
Notes: emp. = employment; ins. emp. = insurable employment; marg. emp. = marginal employment; disp. inc. = disposable income; min. wage = minimum wage; inc = income.

consumer spending amid a severe economic crisis implies "high" consumption spending compared to the overall economic situation, classifying it as an expansive narrative. Similarly, the statement "all in all, core inflation will continue to develop only moderately in the forecast period despite a very favorable economic environment overall" (DIW 2015/09) is categorized as a contractive narrative, as forecasters perceive inflation as relatively low considering an overall positive economic environment.

In a few cases, it is not possible to classify a mixed case solely based on forecaster's sentiment, especially when the narrative discusses expansions and contractions in different areas related to the same economic concept. In such instances, I classify the same chunk as both expansive and contractive. This situation commonly arises in relation to taxes, such as when consumption taxes are raised while income taxes are lowered.

Neither expansive nor contractive narratives clearly predominate for GDP or consumption, as Figure 3a and 3b show. Hence, the forecast error signs are likely to occur independently of forecasters' perspectives on growth and consumption. In contrast, correctly predicted positive forecast errors with low loss tend to include more expansive fiscal policy narratives. The most expansive government narratives discuss stimulus to counteract economic recessions. While narratives stating reduced budget deficits or expiring government aid appear for both forecast errors, they predominate in the subsample with overestimated consumption forecasts.

A substantially higher share of chunks discusses overall employment. More than one-third of the subsample chunks with positive forecast errors indicate or expect rising employment. Noticeably, around one-seventh of the chunks specifically mention shrinking marginal employment. Most of these chunks still mention rising overall employment due to higher rises of employment subject to social contribution. Fewer low-loss chunks with negative forecast errors discuss employment. However, if forecasters discuss the labor market in these chunks, they tend to express weak or shrinking employment.

Another common occurrence within positive forecast error chunks are narratives, including wage and transfer payment increases (Figure 3c). Particularly, narratives concerning the introduction or rise of the minimum wage appear in nearly half the chunks with positive forecast error. The German minimum wage was introduced in January 2015 and increased seven times until 2022. While a significant share of the model's best predictions contain minimum wage narratives, the minimum wage subsample scores only slightly higher than the overall corpus, with an MCC of 0.25. Thus, the model only excels in classifying specific narratives concerning minimum wages. The negative forecast error sample contains no minimum wage narratives and overall fewer examinations of wages and transfer payments (Figure 3d).

The last group of narratives relates to prices, exchange rates, and taxes. This group is more present in chunks with overestimated consumption forecasts among the low-loss subsample. BERT particularly excels in identifying chunks with negative forecast errors that contain narratives about rising inflation and taxes. In contrast, forecasters underestimated consumption when they found prices and taxes low or decreasing. Contractive narratives on prices and taxes are also present in chunks with negative forecast errors. However, they often occur in double counted chunks discussing mixed different taxes or price indexes. For example, "although the return of VAT rates to their previous level will lead to an increase in the [tax] rate next year, this will be dampened in particular by the partial abolition of the solidarity surcharge[11] on [NUM]" (IFO 07/2020), counts to both classes.

Most narratives within the low-loss subsample relate to business cycles, prices, and the labor market, confirming the results in Section 4.2. Hence, my findings correspond to Foltas and

11. The solidarity surcharge is a temporary income tax introduced in 1991 to finance the German Unification. In 2021, the tax was abolished for low and medium income households.

Table 4. Narratives: Labor market (I)

Institute	Date	FE	Loss	Text snippet
IWK	09/2014	positive	0.10	the increase in employment will continue in [DATE] table [NUM]. [...] thanks to stable growth, employment will increase again in [DATE]. however, the introduction of the general statutory minimum wage is expected to have a negative effect on employment. [...]
GED	04/2015	positive	0.09	good economic situation conceals employment effects of the minimum wage. [...] in fact, the number of mini-jobs, which has been fairly stable in the past at [NUM] million, has fallen by around [NUM] since october [DATE] on a seasonally adjusted basis. however, the sharp rise in employment subject to social security contributions since november could also be related to the fact that some mini-jobs have been replaced by employment subject to social security contributions. [...] this [price transmission] in itself reduces short-term employment effects, but could have an impact on demand for the corresponding products and lead to a loss of employment in the medium term.
IWK	04/2018	positive	0.10	[...] the high increase in employment once again corresponds to a strong growth in employment subject to social security contributions. this even increased by [NUM] on average over the year [DATE]. in contrast, employment not subject to social security contributions lost significance. for example, exclusively marginal employment fell by [NUM], which may also be an after-effect of the introduction of the statutory minimum wage. since [DATE], the number of minijobs has fallen by almost [NUM].

Notes: FE = forecast error; loss = 1 - predicted probability of the true class. Square brackets signify numerals, dates or annotations for presentation purposes

Pierdzioch (2022b), which show that employment, wage, inflation, and taxation topic proportions within forecast reports have predictive value for forecast errors. However, the BERT model provides further insights as it reveals specific narratives related to forecast errors within these relevant topics. The subsequent sections examine the nature of this potentially underutilized information in greater detail.

5.2 Case study: Labor market-related narratives

Table 4 presents text snippets of low-loss chunks with underestimated consumption forecasts and typical labor market, particularly employment-related narratives. IWK 09/2014, notifies strong rises in employment while warning that the introduction of the minimum wage in 2015 would negatively impact employment growth. GED 04/2015 finds substantial employment increases despite the recent introduction of the minimum wage, but reiterates their concerns of medium-term employment losses. Furthermore, the forecasters suggest that employers replace minijobs[12] with employment subject to social contributions. Three years after the introduction of the minimum wage, IWK 04/2018 finds a continuation of strong employment growth, particularly jobs subject to social security contributions that rose on the costs of marginal employment.

All three chunks contain comparable narratives, for which the model correctly predicts an underestimation of consumption forecasts with a probability of 85% or higher. The forecasters

12. Minijobs are a form of marginal employment introduced in 2003, which allows employers to pay wages free of income tax and social security contributions up to a specific limit. In 2015, the limit was 450€ per month.

acknowledge a strong labor market while warning about the negative effects on employment from the introduction or increase of the minimum wage. Hence, the underestimation of private consumption could result from an overestimation of the negative labor market impacts of the German minimum wage. Investigations of the German minimum wage's effect on employment yield conflicting results; however, all observed effects, whether negative or positive, are relatively small in size (Bruttel 2019).

Another potential explanation is that the minimum wage and its increases shifts the importance of different forms of employment with impacts on private consumption that are not (fully) reflected in forecasts. Studies suggest the decreased attractiveness of marginal labor forms due to the minimum wage (Bruttel 2019; Holtemöller and Pohle 2020). Employees in marginal employment often receive low household income, which is associated with a higher than average propensity to consume (Dynan 2012); hence, employees transitioning into better-paid forms of labor could positively impact consumption. If such an effect exists, forecasters seem to have neglected or underestimated it, as the recurring narrative within the low-loss sample suggests.

Table A1 shows three more chunks with recurring labor market-related narratives with a focus on wages. All chunks observe or expect significantly rising wages, which forecasters attribute to the introduction or the increase of the minimum wage. Additionally, IMK 10/2014 expects rising pensions. GED 10/2015 and DIW 03/2016 find that collectively agreed wages rose less strongly than overall wages. DIW 03/2016 describes the previous acceleration of wage increases as unexpected due to the lower collectively agreed wages. However, they reiterate their forecast of a slowdown in wage increases.

The forecasters' narratives clearly show their expectations of rising wages due to the minimum wage, which was indeed substantial (Bruttel 2019). In particular, wages at the bottom of the scale without collective agreements rose in the years after 2015. Prior to 2015, these wages often stayed behind the overall wage development. As previously mentioned, low-income groups have higher consumption propensities (Dynan 2012); therefore, wage increases within these groups might lead to exceptionally high increases in consumption expenditure, which forecasters might have not fully account for. Remarkably, Bondt, Gieseck, and Zekaite (2020) and Bondt *et al.* (2021) find that disaggregating disposable income into labor, capital, and transfer income to account for different propensities to consume improves private consumption forecasts. The narratives within the low-loss sample suggest that subdividing labor income by employment type or into income brackets could further enhance consumption forecasting.

Table A2 presents typical labor market narratives of chunks with negative forecast error and low loss. Although they tend to have favorable perspectives on the upcoming labor market developments, they describe overall negative conditions. IFW 12/1999 expects that low-wage settlements will stop the reduction of employment. DIW 07/2003 narrative resembles the older forecast while noting extensive transfer payment cuts. One year later, DIW 07/2004 predicts a short-term stagnation in employment. Furthermore, the forecasters note the reduction in regular employment and an increase in self-employed persons.

BERT successfully classifies chunks of overestimated consumption forecasts that contain narratives opposing those of Table 4 and A1. Thereby, contrasting narratives emerge of opposing economic conditions, as the late 90s and early 2000s were characterized by high unemployment and declining real wages (Brenke 2009). While the narratives reflect vastly different economic conditions, the forecasters fail to integrate their information into consumption forecasts efficiently. All narratives predict a turning point, partly due to the wage dynamic. Hence, they resemble their pessimistic counterparts, who predict a slowdown in employment due to rising (minimum) wages.

5.3 Case study: Prices, taxes and government stimulus

Table A3 presents typical price-related narratives with positive forecast errors. IFO 12/2009 finds rising disposable incomes despite a severe recession. Furthermore, falling energy and food prices support consumer budgets. One year later, IFO 12/2010, finds that incomes continued to rise while inflation remained moderate. HWI 06/2013 notify a "deterioration on the labor market" during the European debt crisis combined with an inflation rate well below the 2% target of the EZB. DIW 09/2015 forecasts favorable economic conditions solely impaired by price increases due to rising wages. However, the forecasters notify that a significant fall in energy prices relieves price pressure.

BERT correctly finds all chunks to underestimate consumption. Like the labor market narratives, they contain a balanced consideration of positive and negative economic developments. Among the negative factors are economic weakness and price increases induced by rising labor costs. Positive factors include government stimulus, moderate inflation, and falling energy and food prices. Particularly often, chunks with positive forecast errors and no or only indirect references to the labor market discuss consumer-oriented government responses during crisis years.

Table 5 presents price and tax-related narratives from business cycle reports that overestimate consumption. IFW 12/2009 reports almost constant prices except for rapidly rising oil prices and increases in energy taxes. While DIW 01/2002 assesses the institute's previous year forecasts critically, BERT correctly predicts further underestimation of consumption as they report real income impacts of rising food and energy prices. IFW 12/2006 predicts a continuation of solid expansion, slightly hampered by increased value-added tax. Lastly, IFW 06/2021 finds exceptional strong consumer price increases during recovery due to increased value-added tax, the introduction of a CO2 tax, and recovering energy prices.

Negative consumption narratives with focus on prices contrast their counterparts with underestimated consumption. When forecasters describe the overall economic situation, they use rather favorable terms or neutral terms while notifying a negative impact of consumer price increases. The causes for consumer price increases vary, including rising food and oil costs, increasing consumption taxes, economic recoveries, supply chain issues, and currency fluctuations. Within the low-loss chunks, no chunk with negative forecast error discusses the impacts of price increases due to rising labor costs, as these chunks tend to have positive forecast errors.

6 Conclusion

I have extended previous research on the efficient utilization of economic narratives by macroeconomic forecasters using an adapted and fine-tuned BERT model. The language model has comprehensively quantified textual information into context-aware word representations, fully capturing economic narratives as sense-making stories. Therefore, I have extended previous studies that focus on inefficient incorporation of sentiment or latent topics, which are components rather than economic narratives themselves.

Using the BERT model, I have investigated the consumption forecasts from six German institutes from 1998 to 2021. The model's predictions for a textual paragraph's associated sign of forecast error correlate significantly with its true sign; therefore, I find the forecasters to inefficiently utilize their textual information under flexible loss.

By investigating a subsample of these high-probability textual inputs, I find multiple specific recurring narratives that the model links to the direction of consumption forecast errors. The findings suggest that forecasters do not efficiently transmit the information in narratives concerning employment, wages, transfer income, crisis-related fiscal stimulus, inflation, and consumption taxes into consumption forecasts. Hence, BERT links narratives implying positive net disposable income developments due to one of these factors to consumption underestimations and vice versa. Similar to the results of Stekler and Symington (2016), these findings suggest that forecasters are aware of factors that affect private consumption, while they fail to fully incorporate the information

Table 5. Narratives: Prices, taxes and government stimulus (II)

Institute	Date	FE	Loss	Text snippet
IFW	12/1999	negative	0.07	[...] after a slight decline in the index at the beginning of the year, consumer prices rose significantly from april onwards. this was mainly due to the increase in energy taxes as part of the ecological tax reform and, in particular, the rapid rise in prices on the international crude oil market. excluding these two influences, the price level remained almost constant. [...]
DIW	01/2002	negative	0.06	rarely have assessments of the economic situation in the spring of a year been so far removed from the reality that became apparent a few months later. there were indeed negative developments that could not have been foreseen. these included the price-driving effects of animal diseases as well as rising oil prices, as a result of which real incomes rose noticeably less than was to be expected due to the tangible tax relief. [...] this explains part of the incorrect forecast, but by no means all of it.
IFW	12/2006	negative	0.08	all in all, domestic demand will continue to expand strongly in [DATE], albeit not as quickly as in [DATE]. the dampening effect of the increase in value added tax will be partially offset in [DATE] by the increase in corporate investment activity, which is also partially stimulated by fiscal policy. in the year [DATE], the stimulus from corporate investment will disappear, but the consumer economy will pick up; the expansion will also affect tax policy. [...]
IFW	06/2021	negative	0.01	the strongest rise in consumer prices in [NUM] years is expected this year, with an inflation rate of [NUM]. although this is largely due to special effects such as the increase in value added tax, the introduction of the co [NUM] tax and base effects in the wake of recovering energy prices, the strong recovery in demand following the end of the pandemic is also driving up prices. [...]

Notes: FE = forecast error; loss = 1 - predicted probability of the true class. Square brackets signify numerals, dates or annotations for presentation purposes.

embedded in these narratives into their quantitative assessments of the subsequent dynamics of private consumption.

Particularly, narratives implying changes in net disposable income of low-income groups seem inefficient. This includes labor market narratives indicating wage raises among low-income households and transitions from marginal to regular employment. Furthermore, consumption tax, transfer income cuts, and food and energy prices disproportionately affect the buying power within these groups. Further investigating income effects for different income types (Bondt, Gieseck, and Zekaite 2020; Bondt *et al.* 2021) might provide valuable insights and improve upcoming consumption forecasts.

Another potential explanation for inefficient forecasts lies in the overweighting of adverse factors and the misprediction of turning points. Text samples with positive forecast errors tend to report generally good labor market conditions but include warnings of negative employment effects due to the introduction or increase of minimum wages. In contrast, overestimated forecasts argue that moderate wage settlements will improve the weak labor market. A similar ambivalence is found in narratives on government stimulus, taxes, and inflation. In such a case, inefficiencies arise from forecasters underestimating the persistence of aggregate consumption.

Overall, I find quantifying forecasting narratives with BERT to be a powerful approach for exploring potentially underutilized information in macroeconomic forecasting and a valuable addition to traditional quantitative forecast efficiency analysis. The strength of this approach lies in its capability to reveal circumstances connected to forecast errors that are not captured by traditional indicators. Thereby, the narrative analysis indicates where additional quantitative data could enhance economic forecasting. Future research could investigate the efficient incorporation of narratives in forecasts of other macroeconomic aggregates, countries, or institutions.

Data availability statement

The data that support the findings of this study are available from the author upon reasonable request.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, I used DeepL and Grammarly to improve the clarity and readability of the text. After using these tools, I reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Aftan, S., and H. Shah. 2023. "A survey on BERT and its applications." In *The EdTech and the rise of metaverse,* edited by A. Sarirete, 161–166. Piscataway, NJ: IEEE. ISBN: 979-8-3503-0030-7.

Bondt, G. de, A. Gieseck, P. Herrero, and Z. Zekaite. 2021. "Euro area income and wealth effects: Aggregation issues." *Oxford Bulletin of Economics and Statistics* 83 (6): 1454–1474. https://doi.org/10.1111/obes.12444.

Bondt, G. J. de, A. Gieseck, and Z. Zekaite. 2020. "Thick modelling income and wealth effects: a forecast application to euro area private consumption." *Empirical Economics* 58 (1): 257–286.

Breiman, L. 2001. "Random forests." *Machine Learning* 45 (1): 5–32.

Brenke, K. 2009. "Real wages in Germany: Numerous years of decline." *Weekly Report* 28 (5): 193–202.

Bruttel, O. 2019. "The effects of the new statutory minimum wage in Germany: a first assessment of the evidence." *Journal for Labour Market Research* 53 (1). https://doi.org/10.1186/s12651-019-0258-z.

Chicco, D., and G. Jurman. 2020. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21 (1): 6. https://doi.org/10.1186/s12864-019-6413-7.

Clark, H., M. Pinkovskiy, and X. Sala-i-Martin. 2020. "China's GDP growth may be understated." *China Economic Review* 62:101243. https://doi.org/10.1016/j.chieco.2018.10.010.

Clements, M. P., and J. J. Reade. 2020. "Forecasting and forecast narratives: The Bank of England Inflation Reports." *International Journal of Forecasting* 36 (4): 1488–1500. https://doi.org/10.1016/j.ijforecast.2019.08.013.

Deschamps, B., and P. Bianchi. 2012. "An evaluation of Chinese macroeconomic forecasts." *Journal of Chinese Economic and Business Studies* 10 (3): 229–246. https://doi.org/10.1080/14765284.2012.699704.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805.

Döhrn, R., and C. M. Schmidt. 2011. "Information or institution?" *Jahrbücher für Nationalökonomie und Statistik* 231 (1): 9–27. https://doi.org/10.1515/jbnst-2011-0103.

Dong, Q., L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui. 2022. "A survey on in-context learning," arXiv:2301.00234.

Dovern, J., and J. Weisser. 2011. "Accuracy, unbiasedness and efficiency of professional macroeconomic forecasts: An empirical comparison for the G7." *International Journal of Forecasting* 27 (2): 452–465. https://doi.org/10.1016/j.ijforecast.2010.05.016.

Dynan, K. E. 2012. "Is a household debt overhang holding back consumption?" *Brookings Papers on Economic Activity* 43 (1): 299–362. https://doi.org/10.2139/ssrn.2132615.

Eicher, T. S., D. J. Kuenzel, C. Papageorgiou, and C. Christofides. 2019. "Forecasts in times of crises." *International Journal of Forecasting* 35 (3): 1143–1159.

El Anigri, S., M. M. Himmi, and A. Mahmoudi. 2021. "How BERT's dropout fine-tuning affects text classification?" In *Business intelligence,* edited by M. Fakir, M. Baslam, and R. El Ayachi, 416:130–139. Lecture Notes in Business Information Processing. Cham: Springer International Publishing / Imprint Springer.

Falkner, S., A. Klein, and F. Hutter. 2018. "BOHB: Robust and efficient hyperparameter optimization at scale," arXiv:1807.01774.

Fildes, R., and H. Stekler. 2002. "The state of macroeconomic forecasting." *Journal of Macroeconomics* 24 (4): 435–468. https://doi.org/10.1016/S0164-0704(02)00055-1.

Foltas, A. 2022. "Testing investment forecast efficiency with forecasting narratives." *Jahrbücher für Nationalökonomie und Statistik* 242 (2): 191–222.

Foltas, A., and C. Pierdzioch. 2022a. “Business-cycle reports and the efficiency of macroeconomic forecasts for Germany.” *Applied Economic Letters* 29 (10): 867–872.

———. 2022b. “On the efficiency of German growth forecasts: an empirical analysis using quantile random forests and density forecasts.” *Applied Economic Letters* 29 (17): 1644–1653.

Gasparetto, A., M. Marcuzzo, A. Zangari, and A. Albarelli. 2022. “A survey on text classification algorithms: From text to predictions.” *Information* 13 (2): 83. https://doi.org/10.3390/info13020083.

Guo, X., and H. Yu. 2022. “On the domain adaptation and generalization of pretrained language models: A survey,” arXiv:2211.03154.

Heinisch, K., C. Behrens, J. Döpke, A. Foltas, U. Fritsche, T. Köhler, and H. Reichmayr. 2023. “The IWH Forecasting Dashboard: From forecasts to evaluation and comparison.” *Jahrbücher für Nationalökonomie und Statistik,* no. 0, https://www.degruyter.com/document/doi/10.1515/jbnst-2023-0011/html.

Holtemöller, O., and F. Pohle. 2020. “Employment effects of introducing a minimum wage: The case of Germany.” *Economic Modelling* 89:108–121. https://doi.org/10.1016/j.econmod.2019.10.006.

Izsak, P., M. Berchansky, and O. Levy. 2021. “How to train BERT with an academic budget,” arXiv:2104.07705.

Keidel, A. 2001. “China’s GDP expenditure accounts.” *China Economic Review* 12 (4): 355–367. https://doi.org/10.1016/S1043-951X(01)00073-6.

Krüger, J. J., and J. Hoss. 2012. “German business cycle forecasts, asymmetric loss and financial variables.” *Economics Letters* 114 (3): 284–287. https://doi.org/10.1016/j.econlet.2011.11.005.

Kudo, T., and J. Richardson. 2018. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” arXiv:1808.06226.

Min, S., X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. 2022. “Rethinking the role of demonstrations: What makes in-context learning work?,” arXiv:2202.12837.

Minaee, S., N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. 2022. “Deep learning-based text classification.” *ACM Computing Surveys* 54 (3): 1–40. https://doi.org/10.1145/3439726.

Mincer, J., and V. Zarnowitz. 1969. “The evaluation of economic forecasts.” In *Economic Forecasts and Expectations,* edited by J. Mincer, 81–111. New York: National Bureau of Economic Research.

Müller, K. 2022. “German forecasters’ narratives: How informative are German business cycle forecast reports?” *Empirical Economics* 62 (5): 2373–2415.

Ojala, M., and G. C. Garriga. 2010. “Permutation tests for studying classifier performance.” *The Journal of Machine Learning Research* 11 (2010): 1833–1863.

Patton, A. J., and A. Timmermann. 2007. “Testing forecast optimality under unknown loss.” *Journal of the American Statistical Association* 102 (480): 1172–1184. https://doi.org/10.1198/016214506000001176.

Pierdzioch, C., J.-C. Rülke, and P. Tillmann. 2016. “Using forecasts to uncover the loss function of federal open market committee members.” *Macroeconomic Dynamics* 20 (3): 791–818. https://doi.org/10.1017/S1365100514000625.

Roos, M., and M. Reccius. 2024. “Narratives in economics.” *Journal of Economic Surveys* 38 (2): 303–341. https://doi.org/10.1111/joes.12576.

Sharpe, S. A., N. R. Sinha, and C. A. Hollrah. 2023. “The power of narrative sentiment in economic forecasts.” *International Journal of Forecasting* 39 (3): 1097–1121. https://doi.org/10.1016/j.ijforecast.2022.04.008.

Shiller, R. J. 2017. “Narrative economics.” *American Economic Review* 107 (4): 967–1004. https://doi.org/10.1257/aer.107.4.967.

Stekler, H., and H. Symington. 2016. “Evaluating qualitative forecasts: The FOMC minutes, 2006–2010.” *International Journal of Forecasting* 32 (2): 559–570. https://doi.org/10.1016/j.ijforecast.2015.02.003.

Thompson, B., J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn. 2019. “Overcoming catastrophic forgetting during domain adaptation of neural machine translation.” In *Proceedings of the 2019 Conference of the North,* edited by J. Burstein, C. Doran, and T. Solorio, 2062–2068. Stroudsburg, PA, USA: Association for Computational Linguistics.

Tsuchiya, Y. 2021. “Sources of deterioration in forecast accuracy during the global financial crisis,” https://doi.org/10.2139/ssrn.3809887.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. “Attention is all you need.” *Advances in Neural Information Processing Systems* 30.

Wu, H. X. 2007. “The Chinese GDP growth rate puzzle: How fast has the Chinese economy grown?” *Asian Economic Papers* 6 (1): 1–23. https://doi.org/10.1162/asep.2007.6.1.1.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, *et al.* 2023. “A survey of large language models,” arXiv:2303.18223.

Appendix

Table A1. Narratives: Labor market (II)

Institute	Date	FE	Loss	Text snippet
IMK	10/2014	positive	0.10	in the forecast period, gross wages and salaries will increase significantly [DATE] by [NUM] and [DATE] by [NUM], while net wages will rise only slightly less. the considerable increase for [DATE] is also a reflection of the introduction of the statutory minimum wage. monetary social benefits will increase sharply in both years as a result of the expansion of statutory pension insurance benefits. [...]
GED	10/2015	positive	0.10	[...] the rise in collectively agreed wages slowed noticeably in the first half of the year [DATE]. [...] effective hourly wages rose by [NUM] in the first half of the year, significantly more than collectively agreed wages. this is likely to be largely due to the introduction of the nationwide statutory minimum wage at the turn of the year box [NUM].
DIW	03/2016	positive	0.10	wage increases accelerated at the beginning of the year, although the opposite was to be expected after the collective wage agreements. [...] at the beginning of next year, wage increases will accelerate again, partly due to the impending increase in the statutory minimum wage. after that, they will slow down again somewhat, as indicated by the rather moderate wage agreements already in place, which will take effect [DATE]. [...]

Notes: FE = forecast error; loss = 1 - predicted probability of the true class. Square brackets signify numerals, dates or annotations for presentation purposes.

Table A2. Narratives: Labor market (III)

Institute	Date	FE	Loss	Text snippet
IFW	12/1999	negative	0.06	in western germany in particular, we expect that the reduction in employment will soon come to an end due to significantly lower wage settlements than in this year's wage round and a strengthening of the economy. [...] in the new federal states, the decline in employment is likely to slow noticeably, primarily due to stabilization in the bsm and a somewhat slower pace of structural layoffs in the construction industry and the public sector.
DIW	07/2003	negative	0.07	around the turn of the year [DATE] [DATE], the reduction in employment will come to a standstill. collectively agreed incomes will rise slightly less than this year [NUM]. [...] extensive benefit cuts in the area of monetary social benefits from the state, unemployment benefit II, will dampen the development of mass incomes. [...]
DIW	07/2004	negative	0.08	due to the usual economic lag, employment will not increase until around the middle of this year. [...] as the hartz measures lead to a reduction in traditional labor market measures such as bsm and an increase in new forms such as ichag, the number of self-employed persons will increase in the year [DATE]; at the same time, the number of dependent employees will decrease by [NUM] [NUM] persons less than the number of employed persons. [...]

Notes: FE = forecast error; loss = 1 - predicted probability of the true class. Square brackets signify numerals, dates or annotations for presentation purposes.

Table A3. Narratives: Prices, taxes and government stimulus (I)

Institute	Date	FE	Loss	Text snippet
IFO	12/2009	positive	0.08	private consumption continued to receive fiscal support. private consumption increased significantly in the first half of the year [DATE], adjusted for seasonal and calendar effects. this was mainly due to the fact that the real disposable income of private households rose noticeably despite the severe economic crisis. [...] finally, consumer budgets were relieved by falling energy and food prices. [...]
IFO	12/2010	positive	0.09	[...] according to revised official figures, see box on the revision of the ifo institute's june forecast [DATE], real private consumption rose continuously in the first half of the year; in the second half of the year [DATE] it fell noticeably. the main reason for this development was, on the one hand, that real mass incomes, net wages and monetary social benefits increased noticeably at the beginning of the year. [...] the rise in prices remained moderate. [...] the savings rate rose noticeably.
HWI	06/2013	positive	0.14	[...] private consumption and housing construction will continue to support domestic demand. [...] although the situation on the labor market has deteriorated, it remains robust. inflation has slowed considerably and is likely to remain well below the [NUM] stability mark this year and slightly below it next year.
DIW	09/2015	positive	0.09	[...] labour-intensive goods and services are likely to experience a stronger price increase in view of rising wages; however, the significant fall in energy prices box [NUM] should gradually be passed on to consumers through lower production costs and thus dampen price increases. all in all, core inflation will continue to develop only moderately in the forecast period despite a very favourable economic environment overall. [...]

Notes: FE = forecast error; loss = 1 - predicted probability of the true class. Square brackets signify numerals, dates or annotations for presentation purposes.