Natural Language Processing-Driven Use-Cases for Economic Analysis Using Unstructured Data*

Csanád Temesvári – Beáta Horváth – Lívia Réka Ónozó

Economic text data, such as news articles or retail trade item names, are an alternative, feature-rich, high frequency information source that can provide insight into economic trends and generate timelier and more accurate estimates. We trained multiple deep learning models for two distinct research tasks: 1) the creation of a sentiment index derived from the categorisation of financial and economic articles into three sentiment categories; and 2) the classification of retail trade item names into appropriate tariff categories. Our models consistently outperformed their baseline counterparts for retail trade item classification, while our sentiment index was able to accurately predict economic downturns where high-frequency data were not available.

Journal of Economic Literature (JEL) codes: C43, C45, C60

Keywords: Natural Language Processing, Deep Learning, macroeconomic nowcasting, classification

1. Introduction

The rapid growth of Natural Language Processing (NLP) and the availability of large-scale textual data provide a powerful tool for economists to analyse information. Within economics, the text-as-data paradigm has become one of the most dynamic methodological frontiers, allowing researchers to extract sentiment and narratives directly from unstructured sources such as news, corporate filings, and online media. Recent comprehensive reviews, such as *Ash – Hansen* (*2023*) and *Gentzkow et al.* (*2019*) have documented this transformation, emphasising the centrality of text-based measures in modern empirical economics. More broadly, the intersection between machine learning, NLP, and economic forecasting is now recognised as a key research domain, connecting to foundational work on nowcasting and real-time macroeconomic monitoring (*Babii et al. 2021*).

* The papers in this issue contain the views of the authors which are not necessarily the same as the official views of the Magyar Nemzeti Bank.

Csanád Temesvári: Magyar Nemzeti Bank, Analyst. Email: temesvarics@mnb.hu
Beáta Horváth: Magyar Nemzeti Bank, Senior Economic Analyst. Email: horvathbea@mnb.hu
Lívia Réka Ónozó: Magyar Nemzeti Bank, Supervisory Advisor; Budapest University of Technology and Economics, PhD Student. Email: onozol@mnb.hu

The first version of the English manuscript was received on 14 March 2025.

DOI: https://doi.org/10.33893/FER.25.1.27

Parallel progress in NLP itself has been driven by deep neural networks and, most notably, transformer architectures (*Vaswani et al. 2017*). Models such as Bidirectional Encoder Representations from Transformers (BERT, *Devlin et al. 2019*) have set new performance benchmarks across diverse text classification and sentiment analysis tasks, and they are still a competitive class of models in the Large Language Model era (*Rostam – Kertész 2025*). These architectures enable contextual understanding that differs from traditional dictionary-based or bag-of-words methods, which struggle with words having multiple meanings and negation. In economics, this has encouraged a different avenue from handcrafted lexicons (*Tetlock 2007, Loughran – McDonald 2011*) toward contextual embeddings, resulting in richer, more accurate representations of economic narratives (*Nasiopoulos et al. 2025*). The main application of transformer-based models is in transfer learning, where a model is first trained on billions or even trillions of words or subwords known as "tokens" in a semi-supervised setting, with the goal of predicting the next token from an input of several tokens. This encourages the model to learn the pattern of the human language to build a foundational knowledge. Next, the model is trained on a task-specific corpora such as economic or financial texts, to achieve superior predictive performance in a domain-specific context, known as "finetuning". This enables the final model to exploit both the knowledge of human language, and to be able to perform in a domain such as finance (*FinBERT, Huang et al. 2022*) or scientific texts (*SciBERT, Beltagy et al. 2019*).

Sentiment analysis of economic news gives insights into public mood and market trends, without delays (*Ónozó et al. 2024b:1*). The ability to generate high-frequency indicators from textual sources provides policymakers and economists with near-real-time signals that complement lagging official statistics. At the EU level as a whole, *de Bondt – Sun* (*2025*) created a classification system using ChatGPT to assign hawkish or dovish sentiment to monthly global PMI reports. The authors successfully used these scores in a regression setting to improve the accuracy of the euro area GDP nowcast estimates. At the country level, *Kalamara et al.* (*2022*) created a sentiment index from articles from three prominent UK newspapers with both occurrence-count and supervised machine learning methods. Their index, in combination with other metrics, had remarkable predictive power for "proxy" metrics, widely used by British economists and policymakers. Both *Aguilar et al.* (*2021*) and *Sobrino et al.* (*2020*) constructed a sentiment index for the Spanish economy via keyword searching, using seven major news sites and the quarterly reports of the Bank of Espana, respectively. Both indices were found to perform better at nowcasting national GDP and GDP growth than a survey-based proxy measurement.

With a limited pool of readily available economically relevant sentiment data, manual annotation is needed to create a sufficient dataset for finetuning.

However, this annotation requires significant manual labour and collaboration. To counteract this, active learning (AL) has emerged as a way to combine a small amount of manually annotated data with a large surplus of available, unlabelled examples. Active learning selectively identifies the most informative examples for labelling, thus considerably reducing the annotation burden, while maintaining (or increasing) model performance. There are multiple strategies to find the most informant datapoints, ones that choose examples which are semantically similar to previously misclassified sentences (*Jiang et al. 2012*) or ones where our model is not as "confident" in its prediction (*Schröder et al. 2022*) or both (*Chen et al. 2011*). While the annotators are still mostly human annotators, there is an increase in using LLMs both as annotators and as methods to choose from the unlabelled data, such as pruning not promising datapoints or ranking the data for selection. LLMs are also being used as a vehicle for creating new labelled instances (*Xia et al. 2025*). These heuristics have been used effectively for multiple different NLP tasks, demonstrating their power in creating efficient data generation for training machine learning models (*Settles 2011; Zhang et al. 2022*). A notable example for using active learning in NLP for Hungarian data is *Üveges et al.* (*2024*), wherein the authors classify legal documents into law categories using deep learning models. The use of active learning managed to narrow down the data needed to reach baseline accuracy by up to 60 per cent.

This paper contributes to the literature by showcasing two instances of using transformer models for classifying Hungarian natural text data. First, we finetune different BERT models to generate sentiment scores for news articles. The data comes from two prominent Hungarian online news outlets. A topic modelling was chosen to create economically and financially relevant news database. We also used different AL heuristics to assess their efficacy in increasing model accuracy. The annotations were carried out using ChatGPT. The sentiment scores are then aggregated into a high-frequency sentiment index. This index is then evaluated for its predictive capacity and timeliness relative to key macroeconomic indicators, including gross domestic product (GDP), the purchasing manager's index (PMI), and the unemployment rate. For evaluating the predictive performance, we use the Granger causality test (*Granger 1969*) and in certain cases, the Toda-Yamamoto causality test (*Toda – Yamamoto 1995*), which are hypothesis tests to measure whether lagged values of one time series have a statistically significant power to predict a second series. We use dynamic time warping (DTW, *Berndt – Clifford 1994*) to measure the alignment between the sentiment indices and macro variables. Moreover, as text-based proxy indicators and indices are an effective way to predict crisis periods (*Baker et al. 2016*), we used a threshold autoregressive distributes lag (TADL, *Tong 1978*) model to assess whether or not the sentiment indices behave differently during crisis periods (such as the great financial crisis or the Covid-19 pandemic). As an additional point of interest, we performed inference with the

trained BERT models on a subset of the FineWeb dataset, a web crawl dataset from the CommonCrawl Repository.

Our second use-case trained transformer models to classify retail store receipt item names into Combined Nomenclature (CN) categories. The data was acquired from the National Tax and Customs Administrations and consists of receipts from major retailers in the country. We finetuned different BERT models to categorise the item names, where we created two different types of embeddings: one based on word co-occurrences, and one where the pretrained model's tokeniser created the vector representation.

The remainder of this paper is organised as follows. *Sections 2* and *3* showcase the two use-cases we investigated, both including the datasets we used, our methodology for the research, and our results. The conclusion follows in *Section 4*.

2. Sentiment analysis of news articles

2.1. Data

The first text data source was a collection of economic and financial news articles from two prominent Hungarian news portals that we scraped with the consent of the media. In accordance with the agreement between the central bank of Hungary and the news outlets, the publishers' name may not be disclosed: therefore, we refer to them as Medium 1 and Medium 2. The dates of the articles range from 1999 to 2020. To filter out economically and financially relevant articles, we used topic modelling, namely latent Dirichlet allocation (LDA). This method is trained on the whole corpus of Hungarian economic news articles to group similar news together. The model assumes that the articles arise from a mixture of latent "topics", and the number of topics is a hyperparameter. The training of the model starts out with a probability distribution for each article over the topics and each topic over the set of words and continually updates these using the word co-occurrences between different news in the corpus. By the end, each article will have a certain probability to fall into each topic. We used a grid-search approach to tune the hyperparameters, including the number of topics, and the number of articles the model processed in each iteration. The final model was chosen based on its perplexity, which measures the model's ability to generalise to new data. The final model used 16 categories and after a manual inspection of the top 20 most probable words for each topic, 13 of these categories were economically and financially relevant. The final dataset was constructed by filtering out the articles whose most probable topic was not in the selected relevant topics.

Figure 1 depicts the distribution of the number of articles broken down by year of publishing, while *Table 1* breaks down the numbers by medium. Our goal was two-fold: First, to train a machine learning model to assign a sentiment to a given

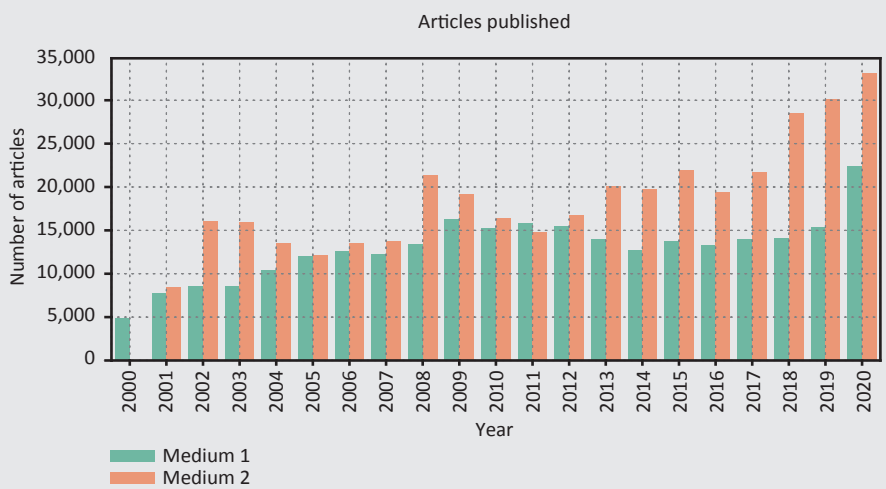
economic or financial article, whether it is positive, neutral or negative. Second, to aggregate these scores into a monthly sentiment index to assess its power to predict the drifts of the economy. To create a small, but usable labelled dataset, out of the 9 million sentences contained in the 700,000 articles we manually annotated 1,645 through a consensus approach, where three economic experts from the Magyar Nemzeti Bank (MNB, the central bank of Hungary) assigned a sentiment to each sentence, and the sentences' label was accepted if two out of the three votes matched. Out of these, a majority vote only occurred for 87 per cent of them, narrowing the dataset down to 1,431. For the training-validation-test data, we split the remaining articles according to a 70–20–10 per cent split, giving us 1,000 articles as training data, 287 as validation data for hyperparameter tuning and 144 for testing.

Table 1
Parameters of the text corpora

	Article count	Average monthly article count	First article	Last article
Medium 1	293,665	1,112	15-02-2000	31-12-2020
Medium 2	404,894	1,533	01-01-2000	31-12-2020
Total	698,545	2,645	01-01-2000	31-12-2020

Source: Arthur et al. (2023:3)

Figure 1
Annual evolution of the number of articles



Source: Authors' calculations

2.2. Methodology

For a baseline, we established a dictionary-based method. This approach uses a set of pre-defined words, which convey either positive or negative economic sentiment. It has many benefits: it is conceptually simple, it is computationally inexpensive, and the results are easily interpretable. However, the construction of this dictionary is not straightforward. Words that have negative sentiment in a common sense might lose this in a financial context, which makes the use of a general lexicon ineffective (*Loughran – McDonald 2011*). This led us to construct our own financial and economic dictionary. We aimed to not use a fixed corpus, as it would have made the dictionary restricted and would be biased to our specific dataset. The inference of this model consists of counting the number of words that are present in the dictionary, then summing their scores (+1 for a positive sentiment, –1 for a negative), then normalising it by the size of the whole dictionary. This results in a net sentiment score for an article, which conveys the overall sentiment: positive, if this number is greater than zero, negative if less than zero, and neutral otherwise. The index is created by averaging out the scores of all the articles in each month.

Our deep learning solution to classifying the news articles is based on two transformer models, both based on huBERT, a BERT model pretrained on the largest Hungarian web corpus (*Nemeskey 2020*). One of the transformer models is finetuned for named entity recognition (NER) (*Yang – Váradi 2023*), while the other is finetuned for sentiment analysis (*Yang – Laki 2021*). Both models are open-source models and are freely available on the Huggingface[1] platform (*NYTK/named-entity-recognition-nerkor-hubert-hungarian, NYTK/sentiment-hts5-hubert-hungarian*). Our model training included a thorough search for the optimal hyperparameters. We used the freely available Optuna[2] software library to streamline the computation, and all the different combinations were evaluated using the loss on the validation set to avoid overfitting, including the batch size, the number of epochs to train the model for, the dropout rate and the weight decay rate. Our experiment revealed that the batch size, the number of sentences that are given to the model at a time, was the most influential parameter, and therefore the most important to tune. This creates a trade-off between better resource management (small amounts of data create less strain on the system) and better generalisation (small amounts of data give less informant gradient for minimising the loss function). We did not modify any other hyperparameters and used the same configuration for both the NER and sentiment scoring models.

[1] https://www.huggingface.co
[2] https://www.optuna.org/

With our limited amount of annotated data, and a large corpus of unlabelled available news articles, using active learning to improve model performance was a suitable approach. Active learning is a human-in-the-loop methodology to increase the generalisation capabilities of a machine learning model. This entails an iterative fashion of training: after the model is trained on the initial training dataset, we select a subset of the unlabelled data which we think are the most useful for the model based on some heuristic, add these to the training data, re-train the model, etc. This process of iteratively labelling new data aims to use the least amount of data needed to train a performant classifier. Our experiments used three different heuristics. As a baseline, we sampled random sentences from our pool of unlabelled sentences, keeping in mind the distribution of the articles per month.

Our first heuristic uses the embedding vectors produced by our transformer models. We compute the embedding vectors for all the unlabelled sentences. Then we filter out the sentences from the test set, where the model predicted negative for a positive sentence, and vice versa. These data points provide the biggest learning opportunity for the model to correctly classify a sentence's sentiment. Next, we searched for all the unlabelled sentences whose *cosine distance* from any of the misclassified sentences is less than 0.00033 but still positive; this threshold was chosen based on the distribution of the different distances. Our final dataset is a sample from these new sentences. This form of active learning aims to provide better context for the model by labelling sentences which are semantically close to ones the model misclassified.

The second heuristic uses the prediction of the neural network to assess its uncertainty about its generalisation. Since we equipped the transformer models such that it is appropriate for classification, the output is a probability distribution over the available sentiment categories, negative, neutral and positive. We used the *entropy* corresponding to the prediction, calculated via *Equation (1):*

$$H(x) = -\sum_{i=1}^{3} p(x_i) \cdot log_2 p(x_i), \quad (1)$$

where $p(x_i)$ denotes the probability for each of the three possible outcomes. The closer to each other these probabilities are, the more uncertain the model is about its predictions. One outcome with a much higher probability conveys confidence about the model's decision. This approach hypothesises that the data which the model is uncertain about lies on the decision boundary of two categories, and therefore labelling those datapoints helps the model to achieve better accuracy. Our investigation revealed that around the entropy of 1.2 the maximum probability value had a noticeable increase. We used this threshold to sample sentences whose entropy exceeds that value. As a final heuristic, we implemented a sampling method which combined both the embeddings of the sentences as the output of the model

through the method called *uncertainty sampling*. We ranked all the unlabelled sentences according to the metric *m*, which was computed via *Equations (2)–(5)*:

$$s_{LC}(x) = 1 - \max_i p(x_i) =: 1 - p(x_{\max}) \tag{2}$$

$$s_{MG}(x) = |p(x_{max}) - p(x_{max-1})| \tag{3}$$

$$D_{avg}(x) = d_{cos}(x, \overline{x_{sen}}) \tag{4}$$

$$m(x) = \left(1 - D_{avg}(x)\right) \cdot \left(0.6 \cdot s_{LC}(x) + 0.4 \cdot s_{MG}(x)\right) \tag{5}$$

where $s_{LC}(x)$ denotes the *least confidence,* which measures how confident the model is in its most probable prediction, $s_{MG}(x)$ measures the *margin* between the first and second most probable answer, while $D_{avg}(x)$ is the cosine distance of the embedding of the sentence from the average embedding $\overline{x_{sen}}$.

For each heuristic, the samples consisted of 5,000 unlabelled sentences; this was deemed to be an appropriate amount of additional training data for the models. The labelling was done by ChatGPT using the official OpenAI API. As economic news articles contain a large proportion of neutral sentences, we reduced all categories of the newly labelled data to the sentiment with the lowest number of sentences, in order to create a balanced dataset for all active learning strategies. The new labelled sentences were added to the original training data, and the model was retrained with the extended dataset. This iteration of picking out new unlabelled sentences and annotating them using ChatGPT was carried out four times in total to produce the final models.

The time series aggregation methodology followed a hierarchical structure from sentence-level to monthly level indices. First, we extracted the sentence-level sentiment. The article-level sentiment index was constructed by summing the sentiment scores of all sentences within each article, providing a comprehensive measure of the overall sentiment tone of the text. The monthly level sentiment indices were computed by taking the arithmetic mean of all article-level sentiment scores within each month, enabling comparison with monthly macroeconomic data.

We compared the monthly sentiment index aggregated from the sentiment scores to three different macroeconomic time series. *Gross domestic product* (GDP)[3] measures the total value of all final goods and services produced within a country's borders over a given time period. It serves as a key indicator of a nation's economic activity and overall economic health. The *unemployment rate*[4] is a survey-based indicator that represents the percentage of the economically active population

[3] Data source: MNB
[4] Data source: Eurostat

who are without work, according to the International Labour Organization (ILO) definition. The *purchasing managers' index* (PMI)[5] is an economic indicator comprised of monthly reports and surveys from private sector manufacturing firms. It reflects business conditions such as production, new orders, employment, and supplier deliveries, with values above 50 indicating expansion and below 50 indicating contraction.

Identification of the macroeconomic variables was done through a systematic empirical approach using rolling window correlation analysis. Specifically, we employed a rolling window to identify the economic indicator that demonstrated meaningful co-movement with our sentiment indices. This data-driven process made it possible that only those macroeconomic variables were included in the final analysis that exhibited similarities with our sentiment measures. This approach allowed us to capture time-varying relationships and identify patterns across different time periods, which aimed to make robust variable selection. For both GDP and the unemployment rate, year-over-year measures are used, as these indicators are more suited to capturing the long-lasting effects conveyed through news dynamics. Due to the data generation process, there may be discrepancies in the temporal alignment, which could affect the interpretation of the results. The monthly GDP estimate is an internal indicator created by the MNB. The unemployment data come from the Hungarian Central Statistical Office (HCSO). For the direct estimation of the monthly unemployment rate, the HCSO uses state space models to estimate monthly employment and unemployment data (*Horváth – Lovics 2023*). The PMI data is published by the Hungarian Association of Logistics, Purchasing and Inventory Management (HALPIM).

2.3. Evaluation

To assess the efficacy of our training, we used several metrics that are well known for their use in machine learning. For classification, we used the weighted precision, recall and F1 score and the balanced accuracy score. These metrics combine the measures for all classes, computing a weighted average of the metrics for each category with the number of elements as the weight, to give an over-encompassing and balanced metric for the capabilities of the model.

For the generated sentiment index, we used two different time series alignment measures. The first one is the Granger causality test (*Granger 1969*), which is a statistical hypothesis test to assess whether or not a series can predict another one, or rather, whether past values of one series contain information that helps to forecast the other. For two time series *X* and *Y*, two autoregressive models are constructed via *Equation (6)* and *(7)*:

[5] Data source: Investing.com [Hungary Manufacturing Purchasing Managers Index (PMI)]

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t \quad (6)$$

and

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t \quad (7)$$

The null hypothesis states that *X* does not Granger-cause *Y*, which is equivalent of the following:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_q = 0 \quad (8)$$

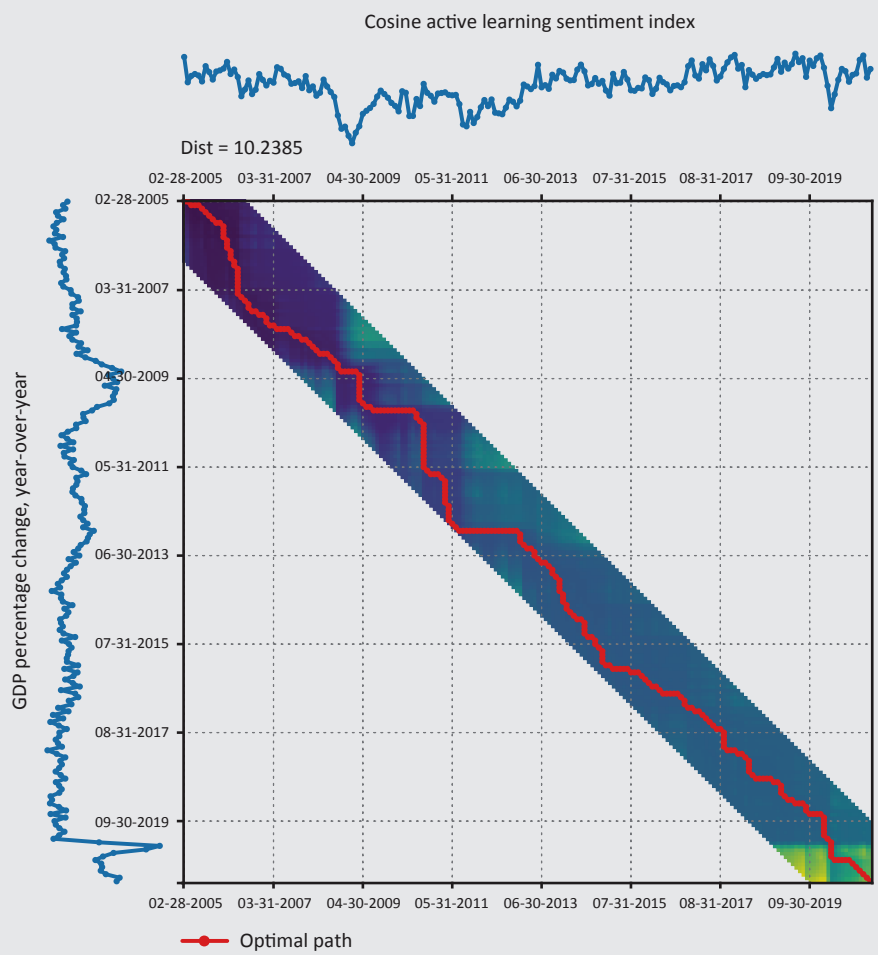
Equation (8) states that the coefficients β_j are all zero, meaning no past values of *X* are significant in predicting *Y*. If we can reject the null hypothesis at a certain level, it means that lagged values of *X* have statistically significant predicting power for future values of *Y*. The Granger causality test assumes that the time series are stationary. This will be tested using the augmented Dickey–Fuller statistical test (ADF, *Dickey – Fuller 1979*). If the results of the ADF test indicate that a time series is non-stationary, the Toda–Yamamoto (*Toda – Yamamoto 1995*) procedure was applied to test for Ganger causality. The method involves determining the maximum order of integration (d_{max}) among the variables and selecting the optimal lag length (*k*) for the standard VAR model. Subsequently, a *VAR(k+*d_{max}*)* model is estimated and the causality test is conducted using Wald tests on the first coefficient matrices, ignoring the last d_{max}. This augmentation enables robust Granger causality testing even when the variables are non-stationary or cointegrated. The order of the autoregressive expressions *p* is a parameter that we need to set ourselves. To find the optimal lag order value, we reported the Bayesian information criterion (BIC, *Schwarz 1978*) and the Hannan–Quinn information criterion (HQ, *Hannan – Quinn 1979*) for all values of *p* up to 8, then chose the one with the smallest BIC and HQ score, and ran the Granger causality test with that parameter.

The second method to evaluate the alignment of two time series is dynamic time warping (DTW, *Proakis – Manolakis 2007*), which tries to find the optimal alignment between two series of numbers. For each point in the first series, the algorithm tries to find the point in the second series with the lowest Euclidean distance to it. The algorithm outputs a pairing, which must satisfy three conditions: 1) each point in one sequence must be paired up with a point in the other; 2) the first and the last points must be paired with each other; and 3) the pairs must be monotonically increasing, meaning the pairs cannot "cross" each other. In practice, another restriction called *global constraint* is also applied, i.e. the admissible points are those whose positions in the two series are "close" to each other. Take, for example a window size greater than 0, then for the coordinate *j* in the first series, the paired up index *i* must be in the interval [*j–w, j+w*]. This not only alleviates a heavy computational burden from trying to compute the distance of every different combination of pairs of points which would result in a quadratic time complexity but

is also in line with a locality argument, namely we are interested in local similarities between our proxy indicators and the macroeconomic variables.

The algorithm produces a pairing of the two time series that pairs the endpoints to each other and is monotonically increasing in both series. *Figure 2* depicts an example of an optimal path between two sequences. The y-axis corresponding to the macroeconomic variable and the x-axis to the sentiment index generated from a model's predictions. The coloured fields are the eligible pairings according to the global constraint, while their colour scale represents the cost of a pairing, measured in the Euclidean distance between the two data points in the series, with darker colour denoting a lower value. The red dots denote the coordinates of the optimal pairings, and the *Dist* field is the sum of all costs of the pairs.

Figure 2
Optimal DTW cost path of the cosine active learning index and the year-over-year GDP index



Source: Authors' calculations

Threshold autoregressive distributed lag (TADL) models were estimated to investigate whether the overall fit between the news indices and macroeconomic data differs during crisis and non-crisis periods. TADL models are particularly useful in capturing regime shifts and nonlinear dynamics in time series data, as they help reveal whether underlying economic relationships change significantly under different conditions. The number of regimes and the threshold values were determined using the Bai–Perron test (see *Bai – Perron 1998*).

The TADL model extends the standard threshold autoregression (TAR) model specification by allowing contemporaneous and distributed lag terms. In the case of a two-regime TADL model, the formulation is shown in *Equation (9)*:

$$y_t = \begin{cases} c_1 + \sum_{i=1}^{p} \alpha_{1,i} y_{t-i} + \sum_{j=0}^{q} \beta_{1,j} x_{t-j} + \varepsilon_t, & \text{if } y_{t-1} \leq \gamma \\ c_2 + \sum_{i=1}^{p} \alpha_{2,i} y_{t-i} + \sum_{j=0}^{q} \beta_{2,j} x_{t-j} + \varepsilon_t, & \text{if } y_{t-1} > \gamma \end{cases}; \quad (9)$$

where y_t is the dependent variable at time t; x_{t-j} are the lagged values of the explanatory variable; y_{t-i} are the lagged values of the dependent variable; $\alpha_{1,i}$ and $\alpha_{2,i}$ are the autoregressive coefficients for the dependent variable in regime 1 and 2, respectively; $\beta_{1,i}$ and $\beta_{2,i}$ are the coefficients for the explanatory variable in each regime; γ is the threshold value that determines the boundary between the two regimes; and ε_t is the error term.

2.4. Results

Table 2 summarises our results for the model training. We chose the sentiment model trained with no active learning as the baseline, as the active learning models were building upon this model, making it a natural candidate. This also allowed us to evaluate how much performance the different active learning heuristics provide. The overall best performing model was the model with the cosine-based active learning, achieving the highest accuracy. The uncertainty-based active learning model had the best performance in categorising neutral sentences, which we deemed important, as the news articles' sentiment is heavily imbalanced towards neutral sentiment. We chose these two models to generate the sentiment indices from the database of news. The optimal DTW distances are introduced in *Table 7* in the *Appendix*. For the PMI, both active learning models generated a comparable result. The dictionary-based method achieved the lowest distance to year-over-year GDP change, while for the unemployment rate, the BERT models achieved a lower distance than the dictionary-based model.

Table 2
Results of sentiment classification finetuning

Model	Precision (%)	Recall (%)	F1-score (%)	Balanced Accuracy (%)
Baseline sentiment model	62.64	63.19	62.47	62.16
Uncertainty active learning (AL) model	65.76	65.28	65.35	65.14
Cosine active learning (AL) model	70.77	70.14	70.29	70.19

Source: Authors' calculations

The optimal lag length analysis yielded the same results for PMI and GDP, using both the BIC and HQ criteria. For the unemployment rate, we considered the HQ results, taking into account the properties of the unemployment rate series. The results can be found in *Table 6* in the *Appendix*.

The Granger causality analysis indicates that all news sentiment indices contain information that helps to forecast the examined macroeconomic indicators. In the case of the unemployment rate, cosine AL sentiment index and uncertainty AL sentiment index, the Toda-Yamamoto methodology was applied. The results of the stationarity tests are presented in *Table 5* in the *Appendix*, while the Granger causality tests are reported in *Table 3*.

Table 3
Granger causality test results: OLL based on HQ and Granger test type as indicated in Table 5

	Cosine AL sentiment index	Uncertainty AL sentiment index	Dictionary-based sentiment index
GDP YOY	0.0678	0.063	0.0000
PMI	0.0004	0.0002	0.0000
Unemployment rate YOY	0.0255	0.017	0.0000

Note: YOY: Year-over-Year
Source: Authors' calculations

The TADL analysis of macroeconomic variables and sentiment indices supports the findings that the overall fit between the news indices and macroeconomic data differs between crisis and non-crisis periods. In the following analysis, the economic indicators are used as dependent variables and the cosine AL sentiment index is used as the independent variable, representing the best-performing model, as shown above. In the case of the PMI, two regimes were identified. The threshold value is 49.5, which aligns the PMI index's formulation, where a value below 50 indicates contraction. When the values are below the threshold, both lagged and coincident sentiment indices are statistically significant, indicating a leading and predictive effect. By contrast, in the other regime, the relationship suggests contemporaneous co-movement. For GDP, the TADL model identifies three distinct regimes: one

that corresponds to periods of substantial decreases, another that corresponds to periods of "normal" changes, and a third that corresponds to periods of substantial increases. The results show that the lagged variables of the news sentiment indices have different effects across these regimes. In the regime with decreasing GDP, the effects are leading meaning the lagged variables have predictive effect, while in the period of "normal" changes, the impact is not statistically significant. By contrast, during the increasing period, the news index suggests a simultaneous association with GDP. For the unemployment rate, the TADL model identifies four distinct regimes: one for large decreases in unemployment rate changes, two for moderate changes, and one for surge increases in unemployment rate changes. The results suggest that very different processes are taking place in the four regimes. Only during periods of surge increases in the unemployment rate changes do the lagged variables of the news index become significant, indicating a leading and predictive effect. The outcomes of the TADL analysis can be found in *Table 8* in the *Appendix*.

2.5. Inference on the Fineweb2 dataset

As a further evaluation of our active learning-aided transformer models, we sought to generate another index from a different Hungarian dataset. The Fineweb2 dataset is a multi-trillion token natural text dataset from more than 1,000 languages, consolidated from the CommonCrawl open repository of web crawl, to create a clean, multilingual dataset for all NLP tasks. Each datum is a snapshot of the textual information from a website, preprocessed through a special pipeline, which includes deduplication of the website content, filtering for NSFW sites, and fixing encoding issues. For a detailed explanation on the preprocessing, see *Penedo et al.* (*2025*).

We selected the Hungarian portion of this dataset, which consisted of text from 50 million websites, with timestamps ranging from March 2013 to April 2024. We used the LDA model trained on the news articles to infer the latent topics and kept the data which fell into the same pre-defined categories. As processing the whole filtered dataset was not feasible on a reasonable timescale, we decide to take a randomised sample, where we chose 100 datapoints from each day. This resulted in 110,000 unique websites, comprised of 3,665,315 sentences which we deemed enough data to create the index. The inference of the models and the index generation were conducted in the same manner as described in *Section 2.2*. *Figure 4* in the *Appendix* shows the chart of the generated indices for the cosine AL model, compared against the year-over-year GDP percentage change. The sentiment index went through a linear scaling, and thus its minimal and maximal value equals the corresponding minimal and maximal value of the GDP change. The sentiment indices seem to follow the dynamics of GDP, especially at the onset of the Covid-19 pandemic. In response to Covid-19, both GDP and sentiment indices experienced a significant decline. However, the drop in the sentiment indices appears to be more persistent. One possible reason for this is that the news, focused for an extended period on the economic burdens caused by Covid and the long-term effects of the

resulting uncertainties. The continuous flow of negative information, combined with the slow recovery of the economy, may have contributed to the indices remaining at a low level for a prolonged period. However, no clear lead-lag relationship could be found between the two series.

3. Retail trade items classification

3.1. Data

For the trade item classification task, we used an extract of the Online Cash Register, a list of approximately 445,000 unique trade item names provided by the Hungarian National Tax and Customs Administration (NAV). Our goal was to finetune a model to predict the item's Combined Nomenclature (CN) code, which is the European Union's standard for classifying retailed goods and is built on the Harmonised System (HS). The CN code is an eight-digit hierarchical code system, with each two additional codes refining each category. In our dataset, only 53,292 unique trade items had a valid four-digit CN code. Within these categories, we selected 17 of these four-digit categories as our focus, as these categories were retail item categories (as our research focus), and with enough items for a finetuning to be effective.

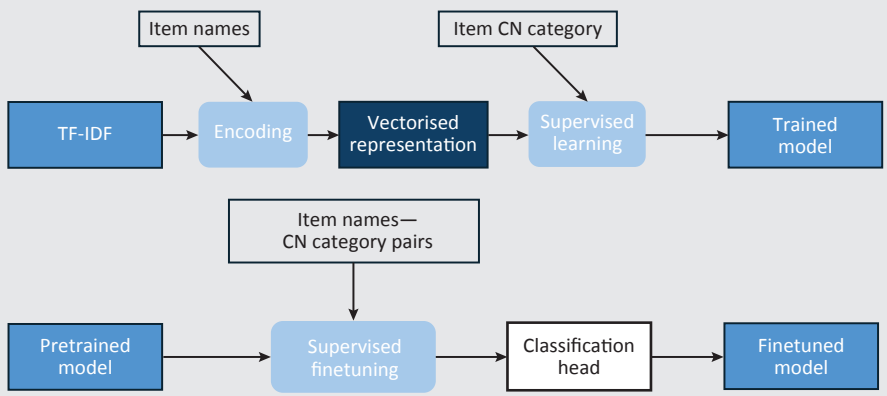
The data cleaning consisted of the following steps: first, cleaning any marker that could point to the retailer the receipt was from (e.g. the name of the store for a store-exclusive product). For privacy and security reasons, this step was carried out by the Tax Authority, which removed any empty items, instances that do not correspond to products (discounts, coupons) and any extra characters (commas, colons, white spaces). Many items had, as part of their name, an indication about the packaging (empties, weight, measurements). For the second step, we created two datasets: one in which this extra information about packaging is removed, and one where it is not, and compared the performance of the models on both datasets. We first split the dataset into an 80–20 per cent part for the train-test split, and then we took a 10-per cent random sample from the training data for the validation split.

3.2. Methodology

We identified two different possibilities to create embeddings. First, we created them through a modified TF-IDF (Term Frequency—Inverse Document Frequency) algorithm (*Sparck Jones 1972*), which uses character co-occurrences to create similar vectors for similar item names. The models trained using the first method are called *vectorised* models. The second option was to let the pretrained model create its own vector representation through its tokeniser. We compared three different models: RoBERTa (*Liu et al. 2019*), huBERT (*Nemeskey 2020*) and PULI (*Yang et al. 2023*). All models are transformer-based open-source models and are available on HuggingFace (*SZTAKI-HLT/huBERT-base-cc, sentence-transformers/all-roberta-large-v1, NYTK/PULI-GPT-2*). For the finetuned models – since language models are next-token predictors – we modified the final layer to align with our goals for

classification. This process led us to create and compare six different models. The final training pipeline of the methodologies is depicted in *Figure 3*.

Figure 3
Schematic of the vectorised learning and finetuning



Source: Authors' compilation

3.3. Results

The results are summarised in *Table 4*. For all models and training, the performance was better on the dataset in which the packaging information was not removed, compared to the one where it was, meaning that this information was helpful for the models to distinguish between different categories. While only looking at the precision values one could think the two different training paradigms were close in performance, a more comprehensive metric such as the balanced accuracy reveals the superiority of finetuning over using only the embeddings generated from the models. This difference indicates that the attention mechanism that drives transformers is capable of capturing semantic relationships between different parts of the items' description better than a simple co-occurrence algorithm. For both vectorised and finetuned models, the huBERT based model was the best performer in most metrics. This is an expected result, as the PULI model is a generative model, which is less suited for classification tasks. Their mutual dominance over the RoBERTa model can be attributed to the fact that, while RoBERTa was pretrained on a multilingual corpus, both huBERT and PULI were pretrained solely on a Hungarian language corpus, making them suitable for tasks with predominantly Hungarian data.

Table 4
Vectorised and finetuned model results

Model	Precision (%)	Recall (%)	F1-score (%)	Balanced Accuracy (%)
Vectorised RoBERTa	61.29	17.32	6.27	6.66
Vectorised huBERT	60.45	17.35	6.34	6.67
Vectorised PULI	61.96	17.22	6.06	6.58
Finetuned RoBERTa	85.84	85.73	85.69	83.42
Finetuned huBERT	88.02	87.97	87.93	85.82
Finetuned PULI	86.63	86.49	86.44	83.95

Source: Authors' calculations

4. Conclusions

Our study investigated Natural Language Processing methodologies, leveraging deep learning models, more precisely transformer-based architectures for creating high-frequency, low-latency indicators from unstructured economic texts. We used financial and economic news articles akin to *Kalamara et al.* (*2022*) and *Aguilar et al.* (*2021*) and retail trade item names and transformed them into actionable classification instruments. These feature-rich sources served as a good basis to generate accurate estimates of economic dynamics, which were validated by the techniques used.

For the sentiment analysis, our deep learning models were effective in assigning sentiment categories to different articles. The use of active learning heuristics enhanced model generalisation capabilities and rendered the models more effective by using less training data needed for finetuning, similar to *Üveges et al.* (*2024*). The cosine-based AL model achieved the highest overall accuracy, while the uncertainty-based AL model demonstrated remarkable performance in classifying neutral sentences, a critical capability given the heavy imbalance toward neutral sentiment in economic news data. The resulting sentiment index proved to be effective in predicting economic downturns, particularly where high-frequency data were unavailable.

The time series analysis revealed complex, regime-dependent predictive relationships between the news sentiment indices and key macroeconomic variables. The main methodologies used in our paper show the following results:

1. Granger causality and dynamic time warping (DTW): The Granger causality analysis indicated that all derived news sentiment indices contain statistically significant information for forecasting the examined macroeconomic indicators.

Regarding optimal alignment, the dictionary-based method yielded the lowest DTW distance when compared to year-over-year GDP change and the PMI, but the active learning-based indices achieved a better fit for unemployment change.

2. Threshold autoregressive distributed lag (TADL) Analysis: The TADL models, estimated to capture regime shifts, confirmed that the correlation between news indices and macroeconomic data differs significantly between crisis and non-crisis periods.

Our results show that in the domain of retail trade item classification, the implemented models – specifically, the finetuned transformer architectures (RoBERTa, huBERT, PULI) – exhibited superior performance relative to their baseline counterparts, and this gain was demonstrated by the balanced accuracy scores. The huBERT model pretrained on a Hungarian language corpus proved to be the optimal performer across most classification models, highlighting the benefit of language-specific pretraining for tasks involving Hungarian language analysis. Furthermore, the inclusion of packaging information within the trade item names enhanced model efficacy, indicating its utility in distinguishing between product categories. A performant classifier in the task of classifying retail items would be a welcome addition to the NLP ecosystem, as misclassifying such items has major legal and material consequences (*Ónozó et al. 2024a:133*).

In summary, this study demonstrates the significant potential of applying advanced deep learning and NLP techniques to unstructured data, offering a valuable toolkit for macroeconomists and decision-makers on the market seeking enhanced predictive power and a deeper understanding of nonlinear economic dynamics. Future research directions could include developing a more sophisticated sentiment category system to create specialised sentiment indices, which would enable more granular analysis of economic narratives and their different impact across economic domains.

References

Aguilar, P. – Ghirelli, C. – Pacce, M. – Urtasun, A. (2021): *Can news help measure economic sentiment? An application in COVID-19 times*. Economics Letters, 199, 109730. https://doi.org/10.1016/j.econlet.2021.109730

Arthur, F.V. – Gyires-Tóth, B. – Debreczeni, M.I. – Ónozó, L.R. (2023): *Language of the Market: NLP-Driven Sentiment Analysis of the Hungarian Economy.* In: 14th IEEE International Conference on Cognitive Infocommunication (CogInfoCom), Budapest, Hungary, pp. 93–98. https://doi.org/10.1109/CogInfoCom59411.2023.10397544

Ash, E. – Hansen, S. (2023): *Text Algorithms in Economics.* Annual Review of Economics, 15: 659–688. https://doi.org/10.1146/annurev-economics-082222-074352

Babii, A. – Ghysels, E. – Striaukas, J. (2021): *Machine Learning Time Series Regressions with an Application to Nowcasting*. Journal of Business & Economic Statistics, 40(3): 1094–1106. https://doi.org/10.1080/07350015.2021.1899933

Bai, J. – Perron, P. (1998): *Estimating and testing linear models with multiple structural changes*. Econometrica, 66(1): 47–78. https://doi.org/10.2307/2998540

Baker, S.R. – Bloom, N. – Davis, S.J. (2016): *Measuring economic policy uncertainty.* Quarterly Journal of Economics, 131(4): 1593–1636. https://doi.org/10.1093/qje/qjw024

Beltagy, I. – Lo, K. – Cohan, A. (2019): *SciBERT: A Pretrained Language Model for Scientific Text*. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Berndt, D.J. – Clifford, J. (1994): *Using dynamic time warping to find patterns in time series.* In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. AAAIWS'94, AAAI Press, Seattle, WA, pp. 359–370. http://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf. Downloaded: 5 May 2025.

Chen, C. – Palmer, A. – Sporleder, C. (2011): *Enhancing active learning for semantic role labeling via compressed dependency trees*. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, pp. 183–191. https://aclanthology.org/I11-1021.pdf. Downloaded: 28 September 2025.

Devlin, J. – Chang, M-W. – Lee, K. – Toutanova, K. (2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1: 4171–4186. https://doi.org/10.18653/v1/N19-1423

De Bondt, G.J. – Sun, Y. (2025): *Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases*. Working Paper 3063, European Central Bank. https://doi.org/10.2866/2788332

Dickey, D.A. – Fuller, W.A. (1979): *Distribution of the estimators for autoregressive time series with a unit root.* Journal of the American Statistical Association, 74(366a): 427–431. https://doi.org/10.1080/01621459.1979.10482531

Granger, C.W.J. (1969): *Investigating causal relations by econometric models and cross-spectral methods.* Econometrica, 37(3): 424–438. https://doi.org/10.2307/1912791

Gentzkow, M. – Kelly, B. – Taddy, M. (2019): *Text as data.* Journal of Economic Literature, 57(3): 535–574. https://doi.org/10.1257/jel.20181020

Hannan, E.J. – Quinn, B.G. (1979): *The determination of the order of an autoregression*. Journal of the Royal Statistical Society: Series B (Methodological), 41(2): 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x

Horváth, B. – Lovics, G. (2023): *Havi munkaügyi adatok becslésének módszertana a KSH-ban* (*Methodology of Monthly Labor Market Data Estimation at the Hungarian Central Statistical Office*). Szigma, 54(3–4): 205–226. https://doi.org/10.15170/SZIGMA.54.1190

Huang, A.H. – Wang, H. – Yang, Y. (2023): *FinBERT: A Large Language Model for Extracting Information from Financial Text*. Contemporary Accounting Research, 40(2): 806–841. https://doi.org/10.1111/1911-3846.12832

Jiang, S. – Pang, G. – Wu, M. – Kuang, L. (2012): *An improved K-nearest-neighbor algorithm for text categorization*. Expert Systems with Applications, 39(1): 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Kalamara, E. – Turrell, A. – Redl, C. – Kapetanios, G. – Kapadia, S. (2022): *Making text count: Economic forecasting using newspaper text*. Journal of Applied Econometrics, 37(5): 896–919. https://doi.org/10.1002/jae.2907

Liu, Y. – Ott, M. – Goyal, N. – Du, J. – Joshi, M. – Chen, D. et al. (2019): *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Loughran, T. – McDonald, B. (2011): *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.* The Journal of Finance, 66(1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Nasiopoulos, D.K. – Roumeliotis, K.I. – Sakas, D.P. – Toudas, K. – Reklitis, P. (2025): *Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models.* International Journal of Financial Studies, 13(2), 75. https://doi.org/10.3390/ijfs13020075

Nemeskey, D.M. (2020): *Natural Language Processing for Language Modeling*. Ph. D. dissertation, Eötvös Loránd University, Budapest. https://doi.org/10.15476/ELTE.2020.066

Ónozó, L.R. – Putz, O. – Járási, I. – Gyires-Tóth, B. (2024a): *Kiskereskedelmi terméknevek kategorizálása Kombinált Nomenklatúra szerint* (*Categorising retail product names according to the Combined Nomenclature*). In: Berend, G. – Gosztolya, G. – Vincze, V. (eds.): XX. Magyar Számítógépes Nyelvészeti Konferencia (XX. Hungarian Computational Linguistics Conference). Szegedi Tudományegyetem, Szeged, Magyarország, pp. 131–144. https://m2.mtmt.hu/gui2/?mode=browse&params=publication;34560678. Downloaded: 3 December 2024.

Ónozó, L.R. – Arthur, F.V. – Gyires-Tóth, B. (2024b): *Leveraging LLMs for Financial News Analysis and Macroeconomic Indicator Nowcasting.* In: IEEE Access, Vol. 12: 160529–160547. https://www.doi.org/10.1109/ACCESS.2024.3488363

Penedo, G. – Kydlíček, H. – Sabolčec, V. – Messmer, B – Foroutan, N. – Kargaran, A.H. et al. (2025): *FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language.* Second Conference on Language Modeling. https://openreview.net/pdf?id=jnRBe6zatP. Downloaded: 13 September 2025.

Proakis, J.G. – Manolakis, D.G. (2007): *Digital Signal Processing: Principles, Algorithms and Applications*, 3rd Edition. Prentice-Hall International, Incorporated. https://uvceee.files.wordpress.com/2016/09/digital_signal_processing_principles_algorithms_and_applications_third_edition.pdf. Downloaded: 28 August 2025.

Rostam, Z.R.K. – Kertész, G. (2025): *Advances in Pre-trained Language Models for Domain-Specific Text Classification: A Systematic Review.* ACM Transactions on Intelligent Systems and Technology, 16(6), 124: 1–41. https://doi.org/10.1145/3763002

Schröder, C. – Niekler, A. – Potthast, M. (2022): *Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers.* In: Muresan, S. – Nakov, P. – Villavicencio, A. (eds.): Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland, pp. 2194–2203. https://doi.org/10.18653/v1/2022.findings-acl.172

Schwarz, G. (1978): *Estimating the Dimension of a Model.* The Annals of Statistics, 6(2): 461–464. http://www.jstor.org/stable/2958889

Settles, B. (2011): *From Theories to Queries: Active Learning in Practice.* In: Guyon, I. – Cawley, G. – Dror, G. – Lemaire, V. – Statnikov, A. (eds.): Active Learning and Experimental Design workshop in conjunction with AISTATS 2010, pp. 1–18. http://proceedings.mlr.press/v16/settles11a/settles11a.pdf. Downloaded: 10 September 2025.

Sobrino, N.D. – Ghirelli, C. – Hurtado, S. – Pérez, J.J. – Urtasun, A. (2020): *The narrative about the economy as a shadow forecast: an analysis using Banco de España quarterly reports.* Working Papers 2042, Banco de España. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/20/Files/dt2042e.pdf

Sparck Jones, K. (1972): *A statistical Interpretation of Term Specificity and its Applications in Retrieval*. Journal of Documentation, 28(1): 11–21. https://doi.org/10.1108/eb026526

Tetlock, P.C. (2007): *Giving content to investor sentiment: The role of media in the stock market*. Journal of Finance, 62(3): 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Toda, H.Y. – Yamamoto, T. (1995): *Statistical inference in vector autoregressions with possibly integrated processes*. Journal of Econometrics, 66(1–2): 225–250. https://doi.org/10.1016/0304-4076(94)01616-8

Tong, H. (1978): *On a Threshold Model in Pattern Recognition and Signal Processing*. In: Chen, C. (ed.): Pattern Recognition and Signal Processing. NATO ASI Series E: Applied Sc., (29). Sijthoff & Noordhoff, Netherlands, pp. 575–586. https://www.researchgate.net/publication/246995827_On_a_Threshold_Model_in_Pattern_Recognition_and_Signal_Processing

Üveges, I. – Vági, R. – Megyeri, A. – Fülöp, A. – Nagy, D. – Vadász, J.P. et al. (2024): *Saving labeling cost by embracing Active Learning: a case study.* In: Berend, G. – Gosztolya, G. – Vincze, V. (eds.): XX. Magyar Számítógépes Nyelvészeti Konferencia (XX. Hungarian Computational Linguistics Conference). Szegedi Tudományegyetem, Szeged, Magyarország, pp. 145–158. https://www.researchgate.net/publication/377730059_Saving_labeling_cost_by_embracing_Active_Learning_a_case_study

Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A.N. et al. (2017): *Attention is All You Need.* Advances in Neural Information Processing Systems 30. Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Downloaded: 4 December 2024.

Xia, Y. – Mukherjee, S. – Xie, Z. – Wu, J. – Li, X. – Aponte, R. et al. (2025): *From Selection to Generation: A Survey of LLM-based Active Learning.* In: Che, W. – Nabende, J. – Shutova, E. – Pilehvar, M.T. (eds.): Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers: 14552–14569. https://doi.org/10.18653/v1/2025.acl-long.708

Yang, Z.G. – Dodé, R. – Ferenczi, G. – Héja, E. – Jelencsik-Mátyus, K. – Kőrös, Á. et al. (2023): *Jönnek a Nagyok! BERT-Large, GPT-2, GPT-3 nyelvmodellek magyar nyelvre* (*The Big Ones are Coming! BERT-Large, GPT-2, GPT-3 language models for Hungarian*). In: 19. Magyar Számítógépes Nyelvészeti Konferencia (19th Hungarian Computational Linguistics Conference), Szegedi Tudományegyetem, Szeged, pp. 247–262. https://acta.bibl.u-szeged.hu/78417/. Downloaded: 12 December 2024.

Yang, Z.G. – Laki, L.J. (2021): *Improving Performance of Sentence-level Sentiment Analysis with Data Augmentation Methods.* In: IEEE (ed.): 12th International Conference on Cognitive Infocommunications (CogInfoCom 2021): Proceedings. Institute of Electrical and Electronics Engineers (IEEE), pp. 417–422.

Yang, Z.G. – Váradi, T. (2023): *Training Experimental Language Models with Low Resources, for the Hungarian Language*. Acta Polytechnica Hungarica, 20(5): 169–188. https://doi.org/10.12700/APH.20.5.2023.5.11

Zhang, Z. – Strubell, E. – Hovy, E. (2022): *A Survey of Active Learning for Natural Language Processing.* In: Goldberg, Y. – Kozareva, Z. – Zhang, Y. (eds.): Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 6166–6190. https://doi.org/10.18653/v1/2022.emnlp-main.414

Appendix

Table 5
Results of the ADF and KPSS stationarity tests

	p-values of ADF	LM statistics for KPSS	Integration	Granger test
GDP YOY	0.0095	0.1483	I(0)	standard
Unemployment YOY	0.5325	0.4619	I(1)	Toda–Yamamoto
PMI	0.0000	0.1962	I(0)	standard
Cosine AL sentiment index	0.0042	0.6134	I(1)	Toda–Yamamoto
Uncertainty AL sentiment index	0.0044	0.8021	I(1)	Toda–Yamamoto
Dictionary-based sentiment index	0.0023	0.2500	I(0)	standard

Note: YOY: Year-over-Year
Source: Authors' calculations

Table 6
Optimal lag lengths for variable pairs based on Bayesian and Hannan-Quinn information criteria

Variable pairs	BIC	HQ	Included Observations
GDP YOY – Dictionary-based sentiment index	1	1	184
GDP YOY – Cosine AL sentiment index	1	1	184
GDP YOY – Uncertainty AL sentiment index	1	1	184
Unemployment YOY – Dictionary-based sentiment index	1	1	172
Unemployment YOY – Cosine AL sentiment index	1	1	172
Unemployment YOY – Uncertainty AL sentiment index	1	1	172
PMI – Dictionary-based sentiment index	1	3	184
PMI-Cosine AL sentiment index	1	3	184
PMI – Uncertainty AL sentiment index	1	4	184

Note: YOY: Year-over-Year
Source: Authors' calculations

Table 7
DTW values for all sentiment indices and macro-variables involved

	Cosine AL sentiment index	Uncertainty AL sentiment index	Dictionary-based sentiment index
GDP YOY	10.2385	10.6563	4.4337
Unemployment YOY	15.6824	16.0024	18.099
PMI	8.4366	8.8464	6.8814

Note: YOY: Year-over-Year
Source: Authors' calculations

Table 8
Results of TADL analysis on GDP, unemployment and PMI statistics using the Cosine AL index

Dependent Variable: GDP YOY Threshold variable: GDP YOY(–1)		
Variable	**Coeff**	**Std. Error**
GDP YOY(–1) < **–1.74** -- 28 obs		
C	–5.39***	0.36
COSINE AL	–195.65***	34.06
COSINE AL(–1)	223.39***	36.10
–1.74 <= GDP YOY(–1) < **2.44** -- 66 obs		
C	0.55**	0.25
COSINE AL	36.10*	20.72
COSINE AL(–1)	6.10	21.12
2.44 <= GDP YOY(–1) -- 97 obs		
C	1.75***	0.47726
COSINE AL	62.51***	17.04485
COSINE AL(–1)	10.22	18.2724

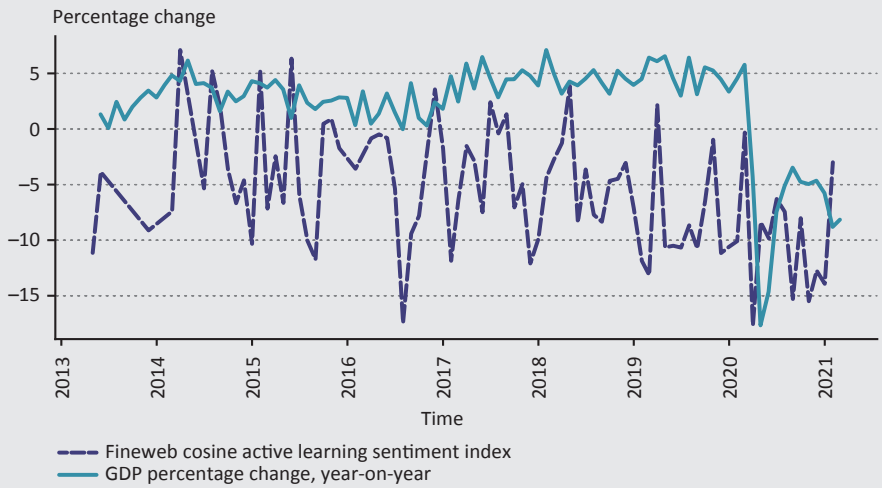
Dependent Variable: Unemployment Threshold variable: Unemployment(–1)		
Variable	**Coeff**	**Std. Error**
Unemployment(–1) < –10.53 -- 56 obs		
C	–26.13***	3.12
COSINE AL	146.06*	87.13
COSINE AL(–3)	131.54	89.80
–10.53 <= Unemployment(–1) < **–2.6** -- 37 obs		
C	–3.73**	1.80
COSINE AL	-138.57	95.55
COSINE AL(–3)	43.02	93.84
–2.6 <= Unemployment(–1) < **10** -- 58 obs		
C	1.81	1.11
COSINE AL	–113.62**	55.37
COSINE AL(–3)	100.81	62.56
10 <= Unemployment(–1) -- 28 obs		
C	21.77***	1.52
COSINE AL	356.83***	83.44
COSINE AL(–3)	–307.90***	78.75

Dependent Variable: PMI Threshold variable: PMI(–1)		
Variable	**Coeff**	**Std. Error**
PMI(–1) < **49.5** -- 29 obs		
C	46.87***	0.64
COSINE AL	–59.70	53.46
COSINE AL(–1)	167.05***	50.45
49.5 <= PMI(–1) -- 162 obs		
C	51.78***	0.42
COSINE AL	70.44***	25.04
COSINE AL(–1)	2.58	26.04

*Note: YOY: Year-over-Year, *** p < 0.01, ** p < 0.05, * p < 0.1*

Source: Authors' calculations

Figure 4
Sentiment index from the Fineweb dataset versus year-over-year GDP change



Source: Authors' calculations