# News Sentiments and Expectations: Text Analysis of Hungarian Economic News

By
Lili Kondrát

Submitted to
Central European University PU
Department of Economics and Business

*In partial fulfillment of the requirements for the degree of Master of Arts in Economic Policy in Global Markets*

Supervisors: Miklós Sebők, László Mátyás

*Vienna, Austria*

2021

# Abstract

The paper examines whether a Sentiment Index based on Hungarian economic news can be helpful to predict economic expectations, currently measured in the form of the Consumer Confidence Index (CCI). The sentiment and outlook of the actors in an economy is of key importance for policymakers, knowing more precisely what the consumers or businesses expect can help in policy choices and could lead to better economic outcomes. The currently available survey-based measure of economic expectations is available monthly, therefore methods that could provide more frequent and timely measurements of sentiment could provide huge benefits. The paper builds a substantial dataset of Hungarian economic news articles from 4 leading news portals and applies text analysis to create a monthly Sentiment Index. The paper then examines whether the Sentiment Index is useful for forecasting the Consumer Confidence Index. The hypothesis is tested by ADL(2,2) model and Granger Causality, and the results suggest that the Sentiment Index is a useful predictor for the CCI, but the results depend on the precision of the sentiment analysis. The findings suggest that policymakers could construct the Sentiment Index on a weekly or even a daily basis and use it to finetune their forecasts about economic expectations. More precise forecasts would enable better policy choices, more suited for the current conditions, therefore the Sentiment Index is a useful tool for nowcasting.

*Keywords: consumer expectations, news sentiment analysis*

## Acknowledgements

I would like to thank for Miklós Sebők and László Mátyás for their support and help during this project. I am also grateful for Ádám Vig and Sebestyén Pap, two fellow students, for their technical help and availability to answer any questions and support me in this journey.

# Table of Contents


1. Introduction ..... 1
   1.1. Relevance ..... 2
   1.2. Background: Consumer Confidence Index ..... 3
2. Literature review ..... 5
   2.1. Online news and economic expectations ..... 5
   2.2. News sentiment analysis ..... 6
3. Methodology ..... 12
   3.1. Data collection ..... 12
   3.2. Description of the dataset, filtering ..... 15
   3.3. Sentiment analysis ..... 20
      3.3.1. Dictionary-based classification ..... 21
      3.3.2. Supervised machine learning approach ..... 24
   3.4. Predictive power of the Sentiment Index ..... 30
      3.4.1. ADL model ..... 30
      3.4.2. Granger causality test ..... 31
      3.4.3. Cointegration ..... 32
   3.5. Discussion ..... 33
4. Policy Implications ..... 35
5. Conclusion ..... 36
6. References ..... 38

Appendix A.: Key-word list ..... 40

Appendix B: Examples for manual coding of articles ..... 42

# List of Tables

Table 1. News outlets with the most traffic based on statistics from March, 2021. .....13

Table 2. List of variables in the dataset.....14

Table 3. Sample of articles in the dataset.....16

Table 4. Results of the ADF test for the (level) time series: CCI and Dictionary-based SI.....23

Table 5. Results of the ADF test for the (differenced) time series: CCI and Dictionary-based SI..23

Table 6. Distribution of categories within the Training set.....25

Table 7. Performance of three Machine Learning Algorithms .....26

Table 8. ADF test for the Supervised Sentiment Index .....29

Table 9. ADL(2,2) coefficients for the supervised machine learning approach .....31

Table 10. Granger Causality test for Equation 1 .....31

Table 11. Cointegration test for the SSI and CCI .....32

# List of Figures

Figure 1. Consumer Confidence Index 2010-2020.....4

Figure 2. Number of articles by month, all four news outlets.....15

Figure 3. Number of articles before and after filtering the dataset.....18

Figure 4. Wordcloud of the most frequent words in the dataset .....19

Figure 5. Dictionary-based Sentiment Index and the Consumer Confidence Index.....22

Figure 6. Plot of the stationary time series: first differences of the Dictionary-based SI and the CCI .....24

Figure 7. Consumer Confidence Index and the Supervised Sentiment Index .....28

Figure 8. Consumer Confidence Index and the Sentiment Index (moving average, n=6) ....28

Figure 9. Plot of the stationary time series: first differences of the CCI and the Supervised Sentiment Index .....29

# 1. Introduction

The paper examines whether a Sentiment Index based on Hungarian economic news can provide useful information for the future behavior of economic expectations.

Expectations of consumers and businesses are of key importance for policymakers, knowing more precisely what the members of an economy expect can help in policy choices and could lead to better economic outcomes. Economic expectations are currently measured by monthly surveys, which are available with a time lag and only in discreet intervals due to their nature. Methods that could provide timely measurements of sentiment could be useful for policymakers for nowcasting. The motivation of this paper is to create a Sentiment Index and analyze if it is helpful in the prediction of economic expectations. Sentiment analysis has been used on English language texts for these purposes, but to the current knowledge of the author this has not been done on Hungarian texts.

The paper builds a substantial dataset of Hungarian economic news articles from 4 leading news portals and applies text analysis methods to create a monthly Sentiment Index. The benchmark for the evaluation of the Sentiment Index is the Consumer Confidence Index, a survey-based expectation metrics used in Hungary. The paper examines whether the Sentiment Index follows similar trends as the Consumer Confidence Index, and whether the Sentiment Index provides useful information to predict expectations. In case the Sentiment Index is useful for this purpose, it would mean that news data could be used in more frequent intervals to measure consumer sentiment or obtain information about the current expectations. The hypothesis is that news sentiments contain information about consumers' expectations, and therefore news sentiment will be a useful predictor of the Consumer Confidence Index.

The hypothesis is tested with Autoregressive Distributed Lag (2,2) models and Granger Causality. The results suggest that precisely measured economic news sentiments are helpful in predicting future behavior of the Consumer Confidence Index.

The paper is structured as the followings. Chapter 2. introduces the literature that focused on news sentiment analysis to explain consumers' expectations. Chapter 3.1. provides a thorough description of the data collection method, Chapter 3.2. describes the obtained database. Chapter 3.3. introduces the sentiment analysis with two different methods, with a dictionary and a supervised machine learning approach. Chapter 3.4. describes the results of the time series analysis and Chapter 3.5. sums up the results. Chapter 4. outlines the policy implications of the findings, and Chapter 5. concludes.

## 1.1. Relevance

News coverage provide the basis for consumers' expectations, therefore news have power over consumer sentiment. Consumers get the facts from the media, but they also have an impression about the current state of the economy based on the tone or volume of news. Understanding the relationship between news sentiments and consumer expectations might help to uncover future consumption patterns. Therefore, there is substantial research about the relation of news sentiment and consumer expectations. The bulk of such research focuses on English texts, and there are less projects that use Hungarian text data and investigate the Hungarian context. The current research aims contribute by examining the predictive power of news sentiments on survey-based expectation measures in Hungary.

Consumer's expectations are relevant information for policymakers and economists, since this soft information can help in policy design. Expectations influence purchasing and saving decisions, and policymakers must consider the probable future behavior of consumers to choose the right

policy measures. The consumer expectations are mostly measured by surveys, but due to their time lag and their discrete availability, alternative methods to measure economic sentiment could provide more accurate or timely information about the state of the economy. Research that uncovers new potential use-cases of news sentiments could help shed light on future economic expectation assessment methods.

The paper aims to answer the question: can economic news sentiment in Hungary be used to predict consumer sentiment? The hypothesis is that economic news influence greatly the expectations, and therefore news sentiment can explain sentiment changes of consumers, measured as the Consumer Confidence Index (published monthly). An underlying hypothesis is that the CCI is a good measure of economic sentiment, and it is the benchmark for measuring the economic expectations. The hypothesis will be tested by running Autoregressive Distributed Lag (2,2) model and by calculating Granger Causality estimates.

## 1.2. Background: Consumer Confidence Index

The research question aims to uncover whether news sentiments are helpful in predicting consumers' economic expectations. The benchmark time series of the analysis is the Consumer Confidence Index, which is shortly introduced in this chapter.

The economic expectations of consumers are traditionally measured by monthly surveys, one such metric is the OECD's Consumer Confidence Index (CCI) for Hungary (OECD, 2021). The survey asks randomly selected households about their financial situation, their upcoming savings and purchasing plans, their impressions about the economy. If the indicator's value is above 100 it signifies a strong consumer confidence, a sentiment which would make them save less and spend more. If the value is under 100, it suggests a pessimistic view where consumers are more prone to save. (OECD, 2021)

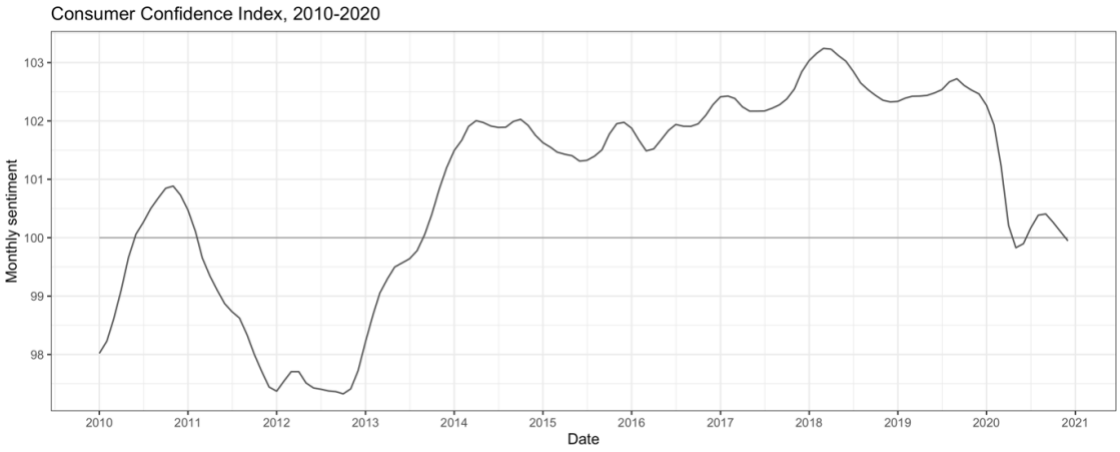


*Figure 1. Consumer Confidence Index 2010-2020, Source: OECD (2021)*

Figure 1. shows the time series of the Consumer Confidence Index for Hungary. The time series shows how the consumer sentiment changed during the recessions in the last decade. The beginning of the series shows the improving sentiment after the financial crisis of 2008-9, and the negative sentiment around 2012-2013 due to the eurozone crisis. The recent coronavirus pandemic also caused a great plunge in the sentiment in 2020.

# 2. Literature review

Chapter 2.1. briefly summarizes the link between online news and economic expectations. Then chapter 2.2 presents previous literature about sentiment analysis of news articles, which provides the basis for the current paper's analysis and describes the evaluation of different text analysis methods in this context.

## 2.1. Online news and economic expectations

Economic expectations of the public matter to policymakers, and its considerations have been present in literature for several decades now. Economic expectations influence economic activity since consumers and businesses alter their spending and saving behavior according to perceived changes in the future. Expectations measured by surveys were shown to have power to predict macroeconomic variables, such as inflation or unemployment (for example: Souleles, 2004; Carroll et al., 1994; Bram and Ludvigson, 1998).

Economic policymakers therefore pay attention to the sentiment of the public, a domain-specific example is the use of forward guidance in central banks. Policymakers benefit from knowing the public's expectations, it allows them to decide about policy measures or fine-tune previously implemented policies. The most common method to collect this information is via surveys, that aim to be nationally representative for a country. However, the survey results might not reflect all available information about economic expectations. In addition, the survey is available monthly, usually with a few weeks lag due to processing time. These drawbacks call for the exploration of new methods to uncover consumer sentiment and provide more timely, frequent, and precise estimates.

One branch of such estimations relies on the sentiment of news articles. The evolving quantitative text analysis tools and sophisticated machine learning models make it possible to process information from vast datasets. This advancement inspired researchers to uncover

sentiment from articles and use that to improve forecasts. The basic mechanism that allows for such analysis is the information mediating role of the media, where the average citizen obtains information.

Media can influence the public through two main channels, as written in Doms and Morin (2004), Ostapenko (2020) and Uhl (2010). One is the mediation of facts and economic variables, providing factual information about the state of the economy. This type of channel is often referred to as 'hard news'. The other channel is through the tone, topics, and extent of news coverage. In practice, the journalists' style, their comments or interpretation about certain events bear relevance, since it influences the impressions of the consumers. This type of information is referred to as 'soft news'. In addition to soft and hard news Doms and Morin (2004) also identifies the volume of news as an important factor in how often consumers update their expectations.

Reading online news articles is a major source of information, both for hard and soft news. In addition to the media, consumers have their own experiences about the economy through other channels as well, for example: number of jobs available in businesses around them, downsizings, prices in the supermarket, wage increases. Therefore, economic expectations are certainly not influenced only by the media, but is serves as a primary resource to supplement personal impressions.



## 2.2. News sentiment analysis

There are two general approaches to conduct sentiment analysis. The first is by using sentiment dictionaries, which contain words or phrases classified as negative, positive or neutral. The texts are searched for these words and the number of words in each category are summed. Then by subtracting the count of positive and negative word occurrences a net sentiment score is created, which can be used for further analysis. This approach is easy and straightforward if a

suitable dictionary is available for the text, but this is not always the case, especially with foreign language texts or domain specific language. The second approach is done by supervised machine learning. In this case a sample of the articles is categorized manually by humans, and the categorization is often verified by several other coders. This sample is called the training set. After manual classification, machine learning algorithms are used to classify the remaining texts (called the test set), by finding hidden patterns in the training set and applying this model for the test set. This approach relies on manual coders and therefore very resource intensive. However, this method can be useful to classify specific texts with more success than the first approach.

The literature that builds on news sentiments is expanding quickly, but most of the research focuses on English corpora. News sentiment analysis is widely used in the Finance literature (for example in Engelberg and Parsons, 2011; Fang and Peress,2009; Griffin, Hirschey, and Kelly, 2011 and Tetlock, 2007), and another branch of literature used sentiment to improve forecasts (for example Garz, 2013 uses it to forecast unemployment expectations, Starr, 2012 investigates whether news shocks have aggregate effects).

The most similar studies to this research question are ones that investigate the relationship of macroeconomic variables and news sentiment (such as Rambaccussing and Kwiatkowski (2020) and Fraiberger (2016), Sharpe, Sinha and Hollrah (2017), Larsen and Thorsrud (2019) and Garcia (2013)), and ones that compare survey results about economic expectations and news sentiment (such as Shapiro et al. (2020), Ostapenko (2020), Uhl (2010), Soroka (2006) and Buckman et al. (2020)). The first branch identifies the relationship between the news sentiments and the state of the economy, and findings are useful in understanding how well the media mediates current events. The second branch answers the question how well news article sentiments reflect economic expectations? This paper aims to answer the latter question, but both types of research listed here are relevant to understand the process and confirm the hypothesis. First the literature that links macroeconomic variables and news sentiments are presented.

Rambaccussing and Kwiatkowski (2020) use sentiment analysis on texts published in printed media to forecast macroeconomic variables in the United Kingdom: inflation, unemployment, and output. The research aims to answer whether point forecasts could be enhanced by including the sentiment information obtained from UK news articles. The analysis is conducted on news articles that focus on the British economy and policy questions; the suitable articles are chosen by keyword filtering. The sentiment analysis is conducted by blending a dictionary-based and a supervised machine learning based approach. The articles were categorized manually by five different coders, and only articles that were unanimously classified as positive or negative were used. The classification was improved by using dictionary-based sentiment measures as features. The authors created quantitative Economic Sentiment Indices from the available text data, using the Support Vector Machine algorithm. The results suggest that the sentiment information improves output and employment forecasts, but it is not useful to nowcast or forecast inflation.

Fraiberger (2016) investigates whether GDP forecasts could be enhanced by including news sentiments as predictors. The author uses a database of Economic news articles in 12 countries and conducts the sentiment analysis with the dictionary-based method. The hypothesis of the research is tested by an autoregressive model supplemented by the Sentiment Index. The paper finds that the index is leading indicator of GDP growth, and concludes that sentiment-based forecasts contain information that are left out of currently used professional forecasts. Forecasts error decreased when the Sentiment Index is added to the models, providing better estimates than currently available ones.

Sharpe, Sinha and Hollrah (2017) applies text analysis to examine the description of the Federal Reserve Board Forecasts published in the Greenbook. The authors tested whether the 'Tonality' of the forecasts' text improves forecasts of inflation, unemployment and GDP growth. To measure tonality in the examined texts, a dictionary-based approach was used, and an index was created. The authors used a custom dictionary, by customizing a widely used dictionary (the Harvard

psycho-social dictionary). Then the authors confirm their hypothesis by multivariate regression that the Tonality index improves forecasts of the chosen variables for the next four quarters.

Larsen and Thorsrud (2019) focuses on narratives (short and easily described events) that matter for business cycles and investigates whether these narratives are associated with economic fundamentals. The main hypothesis is that news that are widely presented in mass media are the most informative about the economy's state. The authors create a quantitative measure by identifying topics with Latent Dirichlet Allocation (LDA) model, which is an unsupervised machine learning model. The authors also use dictionary-based measures in their analysis. The paper concludes that narratives are informative about economic fluctuations and are relevant for high-frequency monitoring of business cycles.

Garcia (2013) focused on asymmetries of the identified relationship between news sentiment and financial variables and hypothesized that the business cycle influences this relationship. The author used sentiment analysis of a New York Times corpus to investigate the effect of news sentiment on asset prices. The main result was that the sentiment information was a useful predictor during recessions, and less so during an economic upswing. The additional research suggests that the results may not be equally strong during the whole timeframe.

In the followings the second relevant branch is introduced: prediction of economic expectations based on news sentiments.

Shapiro et al. (2020)'s research question is the closest to this paper's focus, therefore their findings and methods are useful for present analysis. The paper applies several sentiment measuring methods to answer whether the news sentiments are helpful in predicting survey-based consumer expectations, practically conducting a similar nowcasting exercise as this paper. They apply different pre-existing dictionaries, machine learning algorithms, heuristic rules and create a new dictionary as well. They create a monthly index based on the sentiment analysis of the articles and estimate monthly- and newspaper fixed effects. The authors compare the Sentiment Index to the

next months' survey result, and use an autoregressive model. They find that news sentiments that precede the publication of the survey results are strongly predictive for the survey measures.

Buckman et al. (2020) used the methods described in Shapiro et al. (2020) to leverage the findings during the coronavirus pandemic. The authors constructed and utilized a Daily Sentiment Index to assess consumer sentiment in a timely manner during the coronavirus pandemic, where policymakers were keen to understand the real implications of the pandemic.

Uhl (2010) had similar conclusions. The paper was aiming to explain consumer behavior and expectations with news sentiment. The Sentiment Index was created by using a dictionary-based approach, and the author used that to explain the University of Michigan Index of Consumer Sentiment. The author used ARMA models to test if news sentiment affects consumer sentiments. The findings confirmed that the consumers are influenced by news sentiments, and other variables as well (such as prices, income, and interest rates). The paper also investigated whether private consumption can be explained by news sentiment, and it found that consumption can be best explained by the combination of news sentiment and other variables.

Ostapenko (2020) builds on Shapiro and Wilson (2020), and also focuses on economic expectations and news sentiment but includes a topic analysis to differentiate between the most relevant news topics. The paper considers the fact that households consider different information when developing expectations. Interest rate expectations are mostly influenced by news about the economy, inflation expectations are influenced by loan-related news and unemployment expectations are influenced by housing-related articles. The sentiment results are used in a VAR model that also includes macroeconomic variables. By controlling for the macroeconomic variables, the results are more specifically signaling the effect of soft news on expectations. The study finds that soft news shocks can explain 20% of output's variation, which suggests that experts' opinion published have significant power.

Other papers highlighted that not only economic fluctuations matter, but the reaction to economic news based on their positive or negative nature might be asymmetric as well.

Soroka (2006) focuses on the asymmetric reaction to good and bad news, while investigating the relationship of the economy, media coverage and public opinion. The paper highlights that bad news matter more regarding public perceptions and builds on the substantial information asymmetry literature to justify this claim. The paper uses an ADL model to measure the effects of the change of positive news coverage and change of negative news coverage. The paper finds significant asymmetry in how economic agents react to both economic conditions and news and finds that negative swings are magnified in both cases. This is because the media reacts by writing more about negative conditions, and readers consider this negative news flow and the actual economic figures too. The paper concludes that media has a role in influencing public opinions and especially during economic hardship.

Finally, some papers applied news sentiment analysis on Hungarian corpora as well. The relevant Hungarian literature is narrow, mostly due to the apparent barrier of the scarcity of resources for Hungarian text analysis, compared to the wide availability of English text analysis tools. The potential methods to answer these research questions are limited in some cases by the availability of word lexicons or other sophisticated and validated algorithms. Despite these challenges, there are a few examples for Hungarian media sentiment analysis. Hangya and Farkas (2017) compare sentiment analysis methods on social media texts, Molnár et al. (2015) used the news sentiment to forecast share price movements, Tóth (2020) investigates the news sentiments regarding the Migrant Quota Referendum. To the knowledge of the author, this research question has not yet been explored on Hungarian corpus.

# 3. Methodology

The aim of this research is to create a monthly Sentiment Index and use it to predict the survey-based monthly Consumer Confidence Index. In the followings first the data collection method will be described (3.1.), then the constructed database (3.2.). Then the sentiment analysis of the collected news articles is introduced with two different approaches (3.3.), and the tests of the Sentiment Index's predictive power are described (3.4)[1]. Finally, the results are interpreted (3.5.). Throughout Chapter 3 the analysis was conducted on a database built for this paper, the data source of figures and tables is also this dataset.

## 3.1. Data collection

The analysis builds on the text of news articles, published between 2010 January and 2020 December. The hypothesis is that economic articles have the most impact on consumers' economic expectations, but a few subtopics could still be more relevant within the economic articles. Especially articles that report about the general state of the Hungarian economy, economic forecasts, details about an upcoming economic policy, trends of the world economy. Therefore, to gain meaningful text data that could possibly influence consumers' expectations, the dataset was filtered to narrow it down to the most relevant ones.

A dataset was built for the purposes of this paper, consisting of news articles for the selected period. The dataset contains articles from some of the largest online media outlets in Hungary: 24.hu, Index.hu, Origo.hu and Portfolio.hu. The selection of the media outlets was based on the



[1] The R code of the analysis is available on GitHub at the following link: https://github.com/kondratlilli/Thesis

traffic of news outlets, the relevance of the news published there, and the balance of pro-government and independent outlets. The data collection process is highly resource-intensive, which set a limit on the number of news outlets that could be covered in this paper.

To compare the traffic of the outlets, the published results of the Digitális Közönségmérési Tanács (2021) were used, which is an organization that measures the digital audience of these websites. Their results are shown in Table 1: the ranked list of the news outlets with the most traffic, which provided the starting point for the selection of outlets for this paper. The measured showed in Table 1. are common indicators for website traffic: Monthly Total Real Users and Monthly Total Page Views, which count distinct users and views respectively.

**Table 1. News outlets with the most traffic based on statistics from March, 2021.**

| Rank | Domain | Monthly Total Real Users | Monthly Total Page Views |
|---|---|---|---|
| **1** | **24.hu** | **3 841 252** | **99 197 903** |
| **2** | **Index.hu** | **3 838 804** | **222 835 755** |
| **3** | Femina.hu | 3 449 232 | 54 709 431 |
| **4** | **Origo.hu** | **3 276 852** | **131 743 894** |
| **5** | Hvg.hu | 3 260 668 | 60 299 991 |
| **6** | **Portfolio.hu** | **3 217 148** | **56 175 787** |
| **7** | Blikk.hu | 3 105 968 | 68 039 020 |
| **8** | Telex.hu | 2 935 628 | 82 400 848 |

*Source: Digitális Közönségmérési tanács, 2021. Selected news outlets are bolded*

24.hu has the highest traffic regarding the Monthly Total Real Users (MTRU), and Index.hu has the second most traffic measured by MTRU, so I decided to include these outlets. Both these outlets are independent sources, they do not have affiliations with political parties[2]. The third on the list is Femina.hu, however this news outlet does not have an Economy section. For this reason, the news published on Femina.hu are not relevant for the analysis, so I decided to exclude it. The following news outlet is Origo.hu with the 4th most MTRU. It is a pro-government outlet (due to ownership affiliated with Fidesz), the first one of this kind in the list, so the high traffic and its

[2] More specifically, Index was considered an independent news outlet until July 2020. In that month the editorial board was changed and most of the journalists left the news outlet, and the orientation of Index shifted towards a pro-government attitude. The journalists who left Index in July due to these changes created Telex, a new independent news outlet. Telex was not included in the current analysis due to its short lifespan.

orientation justifies its selection into the database. The following two news outlets are Hvg.hu and Portfolio.hu. Hvg.hu is similar to the news outlets on the top of the list: it is an independent news source, and it covers a wide range of issues, one of them being the Economy column. Portfolio.hu is slightly different since it is specialized on Economic issues and does not cover other areas extensively. This specialization makes Portfolio.hu a highly relevant news source, which complements the already chosen 3 outlets (24.hu, Index.hu and Origo.hu) better than Hvg.hu would. For this reason, Hvg.hu was excluded from the dataset, and Portfolio.hu was included. As mentioned before, the data collection is resource intensive and set a limit on the number of news outlets that could be covered.

Once the news outlets were selected, the news articles were collected by web scraping, using Python programming language. Web scraping is a method when the text of an article is extracted from the html code of the webpage, and it is a common automated method to obtain large datasets from online sources. The dataset includes articles only from the Economy section from each article, and one observation is one published article.

First a list of the URLs was obtained from the websites, then the text of the article and other features were downloaded. The list of the obtained features for each article (observation) is shown in Table 2.



**Table 2. List of variables in the dataset**

| Variable Name | Description of the variable |
|---|---|
| **Year** | the year when the article was published |
| **Month** | the month when the article was published |
| **Day** | the day when the article was published |
| **URL** | The URL of the article where it was downloaded from |
| **Source** | Categorical variable that stores the publishing news outlet (Potential values: origo, 24.hu, index, portfolio) |
| **Title** | Title of the news article |
| **Lead** | Summary or lead of the article, when available |
| **Text** | Body text of the news article |
| **Tags** | Tags associated with the article, stored as a list |

## 3.2. Description of the dataset, filtering

The dataset at this point contained 197,384 observations in total. The distribution of articles among the 4 news outlets is shown in Figure 2. The number of articles obtained from the news outlets differed substantially, Origo published the most in the Economy column (around 60,000 articles) while 24.hu the least (around 35,000 articles).

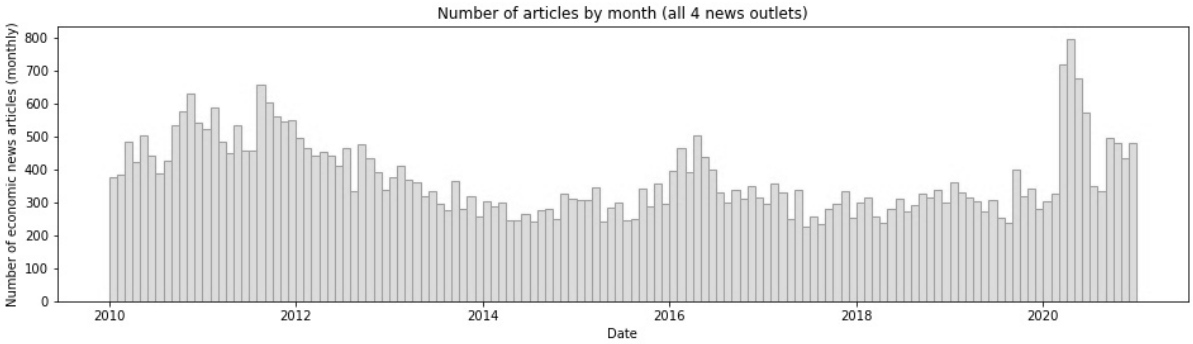


*Figure 2. Number of articles by month, all four news outlets*

The dataset at this point contained every article that was published in the Economy column between 2010-20 at the selected news outlets, in total more than 197,000 articles. As a result, the dataset contained several articles that were not relevant for the analysis, since the Economy column covers a wide range of issues. To showcase the 'noise' in the dataset, Table 3. shows a sample of the articles. The table contains the source of the article, the title of the articles (translated to English) and whether the article is relevant for the analysis. The relevance dummy was hand-coded by the author, based on the topic of the article. The keywords in the relevant articles are bolded. As Table 3. shows, there are topics in the Economy section that are not relevant for the analysis, since they do not inform the reader about the economy or do not influence the readers' economic expectations in any way. Such topics for example are articles that cover political events, articles

about wealth of political figures, guides to personal finances, summary of historical events. The diversity of the dataset requires filtering, to have a subset with only relevant articles.

The data was filtered based on a key-word list. The key-word list was assembled by the author manually. 400 random articles were read and categorized to relevant and not relevant categories, and for each relevant article the keywords were noted.

The most common keywords were kept, and the resulting key-word list contained 84 words, for example: world economy, recession, unemployment, aid, central bank, baby bond, csok. The full list is in Appendix A.

**Table 3. Sample of articles in the dataset**

| | Source | Title (Translated from Hungarian) | Relevant (y/n) |
|---|---|---|---|
| **1** | index | These were the developments of BKK, MÁV and Volánbusz in 2020 | n |
| **2** | index | Asia's richest magnate profited heavily on the coronavirus vaccine | n |
| **3** | index | Bahart is re-announcing the marina | n |
| **4** | 24hu | Orgy, fries, rebellion? Vote for the biggest scandal of 2020! | n |
| **5** | 24hu | The lawn of the Puskás Arena is maintained for 180 million | n |
| **6** | origo | Forty years ago, the U.S. government ran the most peculiar cheese business in history | n |
| **8** | origo | These were the most popular companies among young people in 2020 - gallery | n |
| **9** | origo | We show you the most expensive alcoholic beverages in the world - gallery | n |
| **10** | index | **Refunds** for cancelled bookings are paid to the landlords | y |
| **11** | index | **Oil prices** fell by a fifth this year, no major changes are expected next year | y |
| **12** | index | The Croatian earthquakes caused more than HUF 100 million **damage** in Hungary | y |
| **13** | 24hu | No agreement was reached on the 2021 **minimum wage** | y |
| **14** | origo | Scheduled public sailing **prices** will not change in 2021 | y |
| **15** | origo | **Refunds** for late bookings will be transferred to the landlords | y |

The articles that contained at least one of the words from the key-word list in their title or lead (summary) were kept in the database, the others were dropped. In case of 24.hu many articles did not contain a summary and the keywords were searched only in the title. In case of Origo.hu the leads of the articles were substantially longer than at Index and Portfolio, so the keyword search was conducted on the text of the articles here as well. These changes would expectedly cause more articles to be excluded from the dataset for 24.hu and Origo, but in this case a more narrowed down dataset is preferred.

The filtered database contained 49,502 articles, so roughly 25% of the original observations were kept. The size of the filtered dataset could be easily altered by changing the key-word list, or by looking for the keywords in the body text of the articles. However, in this case a stricter approach was beneficial, which keeps less observations in the dataset. The filtering improved data quality substantially, while the remaining dataset was still large enough to work with.

Figure 3. shows the number of articles before and after filtering the data, by news outlet. The proportion of the articles by news outlet changed substantially. Since Portfolio.hu is specialized on economic news, it was expected that the highest proportion of relevant articles were found among their published pieces: 44% of all the downloaded articles were kept, 56% were dropped. The number of articles decreased most in case of Origo.hu and 24.hu: by 86%, which was expected due to the filtering method. 74% of Index articles were excluded.

The filtered dataset is used for the analysis in the following chapters. To work properly with text data, commonly used preprocessing steps were required, the data cleaning process was conducted as written in Sebők et al., 2021.

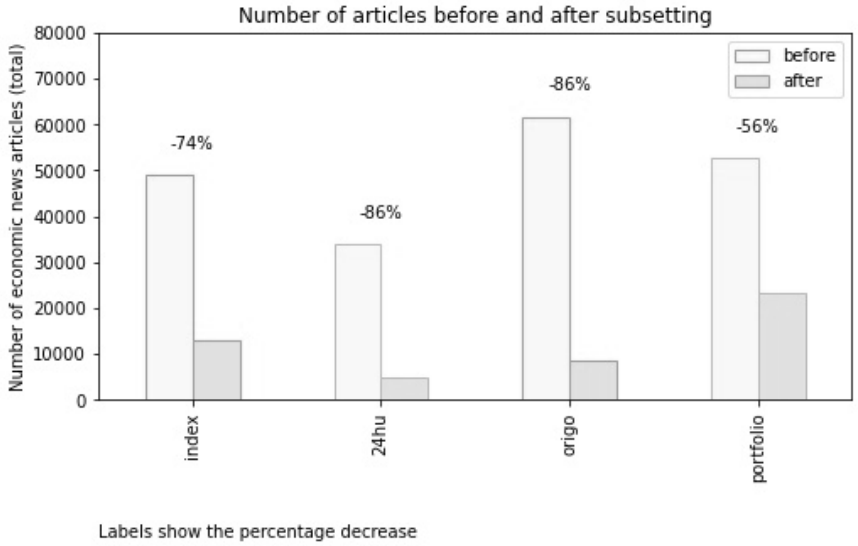


*Figure 3. Number of articles before and after filtering the dataset (reduction from 197,000 to 49,500 articles)*

The article text was tokenized, which breaks down the text to pieces, the result is usually a list of words per document. During tokenization the full text of the article (that includes the title, lead and body text as well) was cleaned of punctuation marks, numbers, and was converted to lower case. The tokens at this point contained stop words, which are frequent words that do not add to the content of the article, such as *the, a, an, but, yet, so, other*. An existing Hungarian stop word list was used (from Sebők et al., 2021) to remove these words, complemented by corpus-specific words that appeared often in news articles, such as *MTI (the Hungarian News Agency), forint, Magyarország (Hungary), vállalat (company), közölt (reported), milliárd (billion).* The most common words in the dataset now are shown in the wordcloud in Figure 4 (in Hungarian language).

After removing the stop words, the tokens were stemmed, which reduced the words to their word stem by chopping the suffixes. This step decreases the number of tokens in the dataset and often allows for more meaningful analysis, since the model would treat words with the same meaning but different suffix as one unit.


*Figure 4. Wordcloud of the most frequent words in the dataset*

After the preprocessing steps, the dataset is converted to a document-term matrix (DTM), which is the common analytical structure to conduct text analysis. Each row in the matrix is a document (article in this case), and each token (word stem in this case) is a column. The values in the matrix reflect the frequency of the word stems in each document. The DTM is a sparse matrix with very high dimensions: it has 49,500 articles as rows, and 230,000 tokens as columns. The size of the matrix can be reduced by excluding rarely used tokens from the analysis. This dimensionality reduction is a beneficial step at this point because it decreases the running time of the code and therefore allows for more efficient analysis. For this reason, the DTM matrix was trimmed and tokens that appeared in less than 15 articles were removed. The resulting trimmed DTM had 49,500 rows and 28,000 columns. The DTM at this point is ready for the analysis.

## 3.3. Sentiment analysis

The research question is whether the sentiment of news articles can predict consumers' economic expectations. To answer the research question, sentiment analysis was conducted on the database of articles, using R programming language. Sentiment analysis aims to categorize text data into positive, negative, and neutral categories (sometimes other categories are identified as well, or only positive and negative categories). The categorization is based on the content, word usage of the articles, mediated mood or opinion, polarity. The sentiment analysis of this paper is a crucial step in the analysis, so two different methods were applied to find the better approach.

Sentiment analysis is often done by using dictionaries: wordlists that contain several hundred words classified as positive and negative. The dictionary-based algorithm searches for the words in the text data and based on the frequency of positive and negative words it calculates a sentiment score. The approach is easy to use and there are ample resources for English language texts. However, the specificity of texts could pose problems, and the dictionary must be well suited for the corpus at hand. The most important limitation for this paper is caused by the language barrier: the available dictionaries for Hungarian language are limited, especially domain-specific dictionaries that contain words frequent in Economic context.

Another approach is the application of machine learning algorithms, commonly supervised learning algorithms. In this case the text data is divided into a training and test set first. Then the articles of the training set are categorized manually to three categories: positive, negative, and neutral. Ideally, the manual coding is verified by more human coders, but in this paper the coding is done solely by the author. When the classification is done and the training set is labelled, a machine learning algorithm is used to fit the model on the data. The algorithm recognizes hidden patterns in train set, and looks for these patterns in the test set. The fitted model then categorizes

the articles of the test set. This approach requires substantial manual coding work, but it is very useful to handle specific texts for which dictionaries are not helpful.

The sentiment analysis was conducted with both methods, the following paragraphs describe the details of the classifications.

### 3.3.1. Dictionary-based classification

To conduct the dictionary-based sentiment analysis, an originally English dictionary was translated and used. A better suited option would be to use a custom-made Hungarian economics-specific dictionary, but this is a very resource-heavy task, so this paper used an already existing dictionary instead. In addition, translation of dictionaries is an imperfect method since the linguistic differences between languages pose challenges, a more precise analysis could be conducted with dictionaries that were made specifically for Hungarian texts.

On the other hand, a translated domain-specific dictionary could still be more useful for the current analysis than a general Hungarian dictionary. The sentiment of Economics texts is hard to evaluate based on general wordlists, so several English domain-specific dictionaries exist. These wordlists include common economic terms. One of the most popular domain-specific dictionaries is the Loughran-McDonald sentiment dictionary, used for the evaluation of financial texts (Loughran, McDonald, 2011). The Loughran-McDonald dictionary identifies 6 categories: positive, negative, constraining, litigious, superfluous and uncertainty. I will only use the positive and negative categories for the analysis. The positive category contains 353 words, the negative category 2354 words. This is a common distribution; positive wordlists are usually shorter in many languages and regarding general and domain-specific dictionaries as well. These wordlists were translated via Google Translate. The translated list contained many duplicates, since in some cases there is only one Hungarian word for several English words. After dropping the duplicates, the length of the Hungarian positive wordlist is 286, while the Hungarian negative wordlist contained 1691 words.

Examples from the positive list are: *bőség (abundance), elismer (acknowledge), siker (success), biztosít (ensure), haszon (gain*); and examples of negative words are: *csőd (bankruptcy), akadály (obstacle), bojkott (boycott), megveszteget (bribe), panasz (complaint), elítél (condemn)*.

The sentiment analysis algorithm in this case simply counts the positive and negative occurrences in each article text and calculates a net sentiment score. After each article was assigned a net sentiment score, the average monthly sentiment score is calculated, this time series will be referred to as the Dictionary-based Sentiment Index (DSI) from now on. The time series of the Dictionary-based Sentiment Index and the Consumer Confidence Index were both normalized, the results are shown in Figure 5.

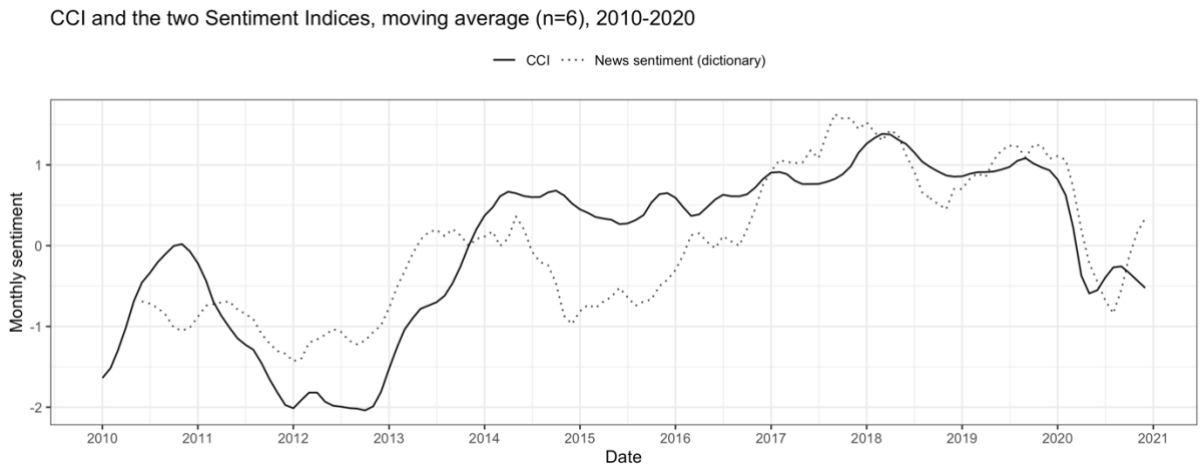


*Figure 5. Dictionary-based Sentiment Index and the Consumer Confidence Index*

The plot shows that in some time periods the results seem reasonable, from 2011Q2 to 2013Q2 and from 2017 onwards the series follow similar trends. However, the Dictionary-based Sentiment Index experienced sharp decreases at the end of 2010 and in 2014, which were not visible in the CCI time series.

To test the hypothesis, an ADL model and Granger causality will be used. To apply these methods, the stationarity of the series will be checked. The stationarity condition is tested by the Augmented Dickey-Fuller t-test (ADF), which tests if the series has a unit root. For both the Consumer Confidence Index and for the Sentiment Index the null hypothesis of nonstationary could not be rejected (results shown in Table 5.)

**Table 4. Results of the ADF test for the (level) time series: CCI and Dictionary-based SI**

| Data: Dictionary-based monthly Sentiment Index | Data: Consumer Confidence Index (monthly) |
|---|---|
| Dickey-Fuller = -2.9064, Lag order = 5, p-value = 0.1996<br>alternative hypothesis: stationary | Dickey-Fuller = -2.3413, Lag order = 5, p-value = 0.4346<br>alternative hypothesis: stationary |

Therefore, the result of the ADF test showed that neither of the time series are stationary, so for the following time series analysis the first differences of the time series will be used. To check the stationarity of the first difference time series, the ADF test was used again. The ADF tests verified that by taking first differences both resulting time series are stationary (results are in Table 6.) The stationary time series are plotted in Figure 6.

**Table 5. Results of the ADF test for the (differenced) time series: CCI and Dictionary-based SI**

| Data: First differences of Dictionary-based monthly Sentiment Index | Data: First differences of Consumer Confidence Index (monthly) |
|---|---|
| Dickey-Fuller = -5.8806, Lag order = 5, p-value = 0.01<br>alternative hypothesis: stationary | Dickey-Fuller = -3.482, Lag order = 5, p-value = 0.04674<br>alternative hypothesis: stationary |

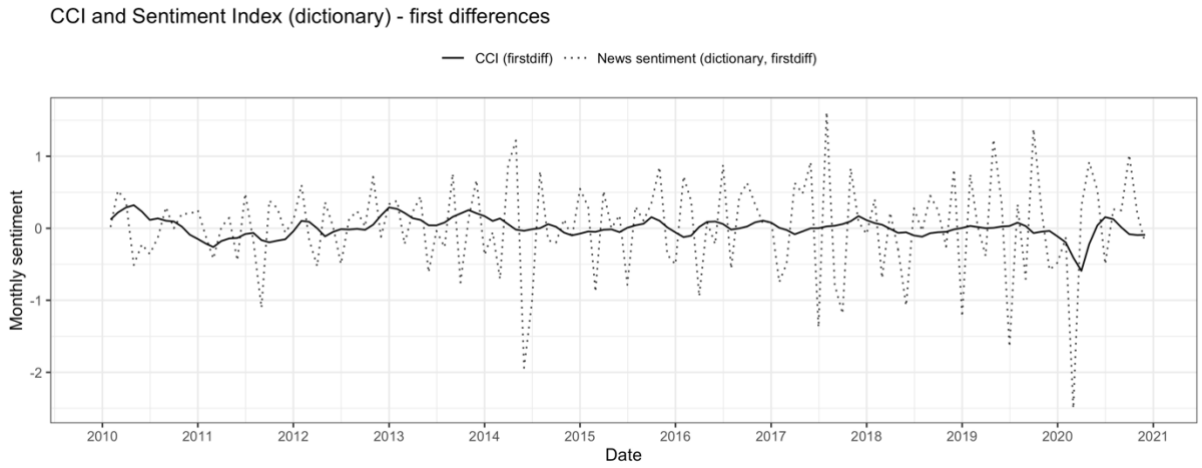


*Figure 6. Plot of the stationary time series: first differences of the Dictionary-based SI and the CCI*

The time series analysis will be conducted in Chapter 3.4., simultaneously comparing the results to the other methods used in this paper.

### 3.3.2. Supervised machine learning approach

The second approach is the supervised machine learning (ML) algorithm. To run the supervised algorithm a labelled (already categorized) train set is required to fit the model. The dataset built from the articles did not contain any labels yet, so manual coding was required to create a proper training set. To have a well-trained algorithm, at least a few hundred labelled observations are needed. I decided to create the training set from 5% of the observations. The total number of observations was 49,502, the size of the training set is 2468 observations. The articles of the training set were randomly selected from the articles, stratified by year, month and the source of the news (news outlet).

The manual categorization was based on the following guidelines. If the article's text is not about a relevant topic and it probably did not influence the readers' expectations, then it is coded as 0 (neutral). If the article is about a relevant topic, but it does not mediate any facts, opinion or

feelings about an event, or it cannot be decided whether the event has positive or negative consequences, then it is coded as 0 as well. If the article is about a relevant topic but mediates both positive and negative opinions in a balanced manner, then it is coded as 0 again (neutral). If the article is relevant and it reports about positive news regarding the economy, then it is coded as 1 (positive). If the article is relevant and it reports about negative news, then it is coded as -1 (negative). The manual classification is a key step in the sentiment analysis process, examples for the above-mentioned cases are presented in Appendix B.

The number of identified positive, negative, and neutral articles were similar, which suggests that the dataset is balanced from the dependent variable's perspective (Table 6.)

**Table 6. Distribution of categories within the Training set**

| **Category** | **Number of articles** |
|---|---|
| Negative articles | 951 (38,5%) |
| Neutral articles | 730 (29,6%) |
| Positive articles | 787 (31,9%) |

In addition to the randomly selected stratified sample, a COVID-19 specific sample was manually coded as well. The number of articles from 2020 is high in the dataset, and the vocabulary of the articles is substantially different than the vocabulary of the articles from 2010-2019. This subsample will be used later to improve the classification by including the COVID-specific training sample.

Once the training sample was available, different machine learning algorithms were fitted and evaluated to find the most suitable model. The model evaluation could be done on labelled datasets only. For this reason, the labelled training set was split further for evaluation purposes: 80% of the observations were used to fit the model and 20% were used as the test set. Based on the performance of the model on the test set, commonly used metrics were calculated: accuracy, recall, precision and the F1 score (written for example in Müller, Guido, 2016).

Accuracy shows the number of classifications that were done correctly divided by the total number of observations (true positives divided by the total number of observations). Precision shows what proportion of the positively categorized observations are classified correctly (number of true positives divided by the sum of true positives and false positives). Recall shows what proportion of all the positives was correctly classified as positive (number of true positives divided by the sum of true positives and false negatives). F1 score is calculated from precision and recall giving an average score about the performance of the model. In the followings the confusion scores are showing the performance of the classification of positive and negative news and exclude the performance of the neutral articles, since these are not considered during the time series analysis and therefore less relevant in this case. (Müller, Guido, 2016)

3 classification methods were tested: Support Vector Machine, Random Forest and the Naïve Bayes model. The confusion scores of the models are shown in Table 7.

**Table 7. Performance of three Machine Learning Algorithms**

| Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Support Vector Machine | 0,6973 | 0,6684 | 0,6635 | 0,6659 |
| Naive Bayes | 0,7384 | 0,7358 | 0,7330 | 0,7344 |
| Random Forest | 0,6923 | 0,6873 | 0,6867 | 0,6870 |

The results suggest that the Naïve Bayes classification provides the best results. However, the method has to be validated on the test set as well. So, I proceeded with the three models and classified the remaining 47,000 observations in the test set. The resulting classification cannot be evaluated by the scores mentioned before, since in the test set there were no labels to compare the results with. Instead, the classification could be validated by taking a random sample from the test set and checking manually whether the classification is correct. In addition, the time series were plotted to visually confirm the results. I validated a random sample of 100 observations and compared the results, since the performance of the models were similar. The results of the manual

validation and the graphs of the time series suggested that even though the Naïve Bayes gave better scores than the other two methods, the Random Forest seemed to return better results, so it was used as the main supervised learning algorithm.

Now the articles are classified into positive, negative, and neutral categories, the Sentiment Index can be constructed. The Sentiment Index is a monthly index calculated from the sentiment calculated in the previous paragraph. The monthly aggregation is necessary to match the time frequency of the Consumer Confidence Index.

The Sentiment Index is calculated as the average net sentiment in that month, from now on this index will be referred to as the Supervised Sentiment Index. The sum of the positive and negative scores equals the net sentiment, which is divided by the number of articles in the database in that month. The resulting Sentiment Index has a smaller range than the Consumer Confidence Index, so the time series will be normalized.

The time series of the constructed Sentiment Index is very noise, with relatively large jumps and decreases between months compared to the changes in the CCI index. This was expected, since the news tend to introduce events dramatically and the data collection and categorization could still leave noise as well. Figure 6. shows the plot of the Consumer Confidence Index and the Supervised Sentiment Index, and Figure 8. shows the moving average of the Supervised Sentiment Index (with n=6).

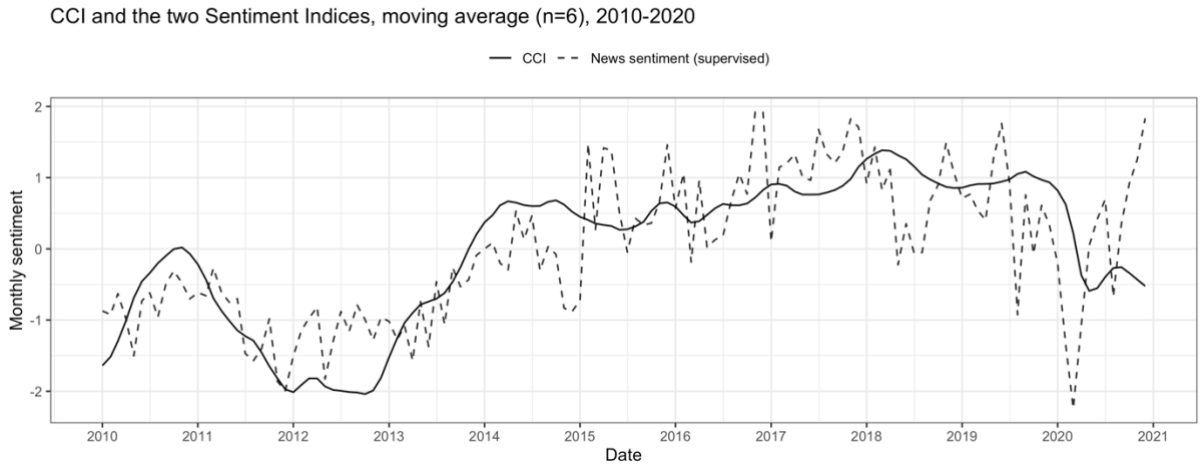


*Figure 7. Consumer Confidence Index and the Supervised Sentiment Index*

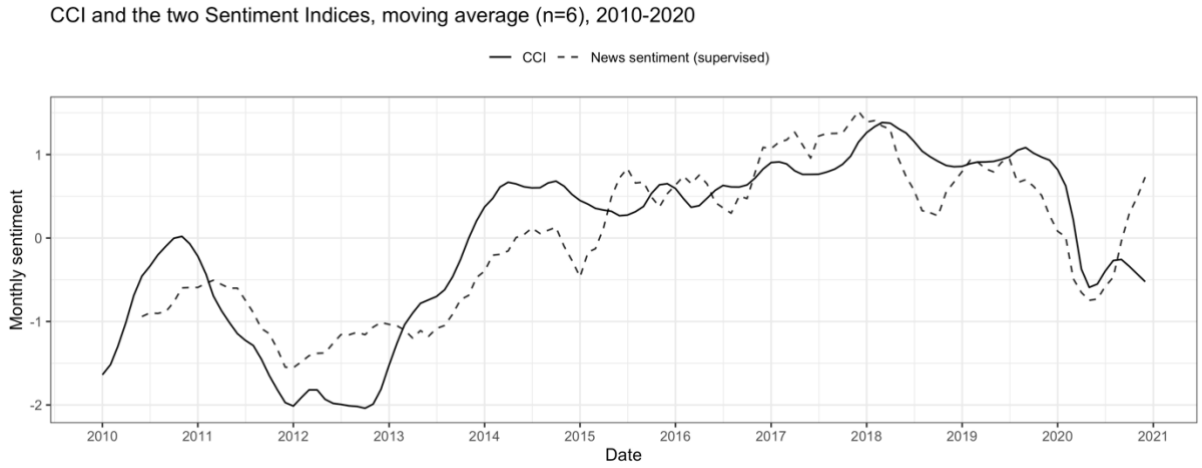


*Figure 8. Consumer Confidence Index and the Sentiment Index (moving average, n=6)*

The plots show that generally the Supervised Sentiment Index performs well, but similar deviations can be noticed as in case of the Dictionary-based Sentiment Index. The Supervised Sentiment Index have similar trends to the Consumer Confidence Index until 2016, but in 2017 the Sentiment Index shows a sharp decline whereas the CCI does not. From 2018 onwards the time series seem to follow similar trends again.

The Augmented Dickey-Fuller test was conducted on the Supervised Sentiment Index, and the results showed that the time series is not stationary. The first difference of the Supervised Sentiment Index was taken, and this time series is used in the following analysis. Table 9. shows the results of the ADF test for the level time series and the first differences of the SI (supervised).

**Table 8. ADF test for the Supervised Sentiment Index**

| data: SI (supervised) | data: First differences of the SI (supervised) |
|---|---|
| Dickey-Fuller = - -2.4949, Lag order = 5, p-value = 0.3707<br>alternative hypothesis: stationary | Dickey-Fuller = -6.5985, Lag order = 5, p-value = 0.01<br>alternative hypothesis: stationary |

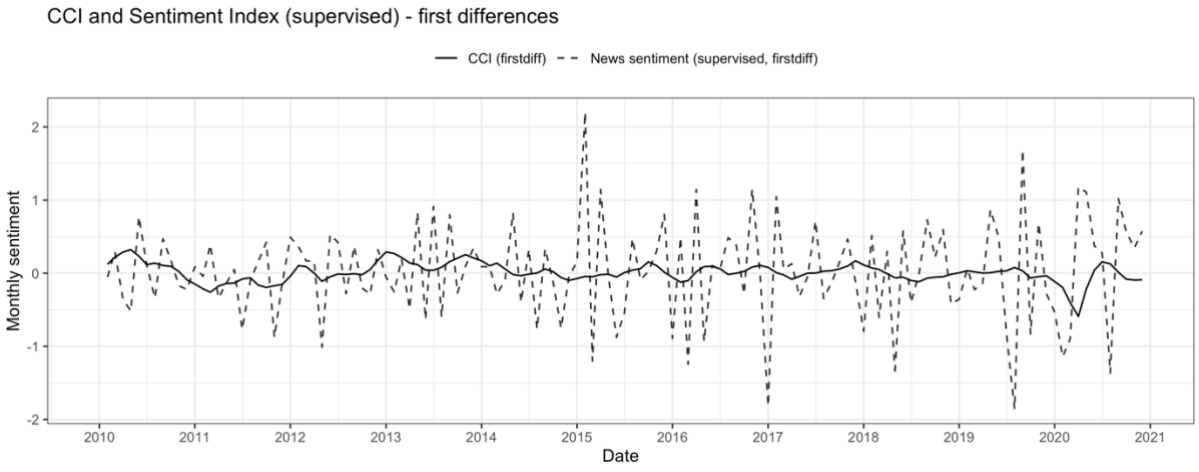


*Figure 9. Plot of the stationary time series: first differences of the CCI and the Supervised Sentiment Index*

## 3.4. Predictive power of the Sentiment Index

The research's hypothesis is be tested by two methods: first running an ADL model on the Consumer Confidence Index and the Sentiment Index, then calculating Granger causality. The tests were conducted based on Brooks (2008) and Hanck (2019).

### 3.4.1. ADL model

The Autoregressive Distributed Lag model has the Consumer Confidence Index at time t as the dependent variable and have lags of the CCI and of the SI on the right-hand side as predictors. The predictive power of consumer sentiment was tested with this model in papers with similar research question, such as Shapiro et. al (2020) and Soroka (2006). The lag length in the model was decided by the Bayes Information Criterion, and the optimal lag length was identified as 2 (based on Hanck, 2019).

The equation is specified such that 2 lags of the CCI and 2 lags of the SI are included, so an ADL (2,2) model is estimated. The model was estimated first with the dictionary-based Sentiment Index, and then with the supervised machine learning-based Sentiment Index. The equation of the estimation is as follows:

$$\Delta CCI_t = \alpha + \beta_1 \Delta CCI_{t-1} + \beta_2 \Delta CCI_{t-2} + \beta_3 \Delta SI_{t-1} + \beta_3 \Delta SI_{t-2}$$

*Equation 1.*



The coefficients of the Dictionary-based Sentiment Index were not significant, so their results are not included in this paper. The Supervised Sentiment Index performed better, the second lag of the Supervised SI is significant, shown in Table 9.

**Table 9. ADL(2,2) coefficients for the supervised machine learning approach**

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 0,0003 | 0,0056 | 0,0462 | 0,9632 | |
| $\Delta CCI_{t-1}$ | 1,2278 | 0,1480 | 8,2979 | 0,0000 | *** |
| $\Delta CCI_{t-2}$ | -0,4802 | 0,0796 | -6,0300 | 0,0000 | *** |
| $\Delta SI(supervised)_{t-1}$ | 0,0169 | 0,0114 | 1,4886 | 0,1392 | |
| $\Delta SI(supervised)_{t-2}$ | 0,0153 | 0,0080 | 1,9216 | 0,0570 | . |

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The results show that the second lagged coefficient of the Supervised Sentiment Index is significant at 10%. These results suggest that news sentiment is indeed a helpful predictor of consumer expectations, but the precision of measuring news sentiments matter greatly.

## 3.4.2. Granger causality test

The Granger Causality test was conducted for the same equation (based on Hanck, 2019). The results were significant for the Supervised Sentiment Index, while not significant for the dictionary-based Sentiment Index. The results for the Supervised Sentiment Index are shown in Table 10.

**Table 10. Granger Causality test for Equation 1**

Model 1: *(Equation 1.) with Supervised Sentiment Index*

Model 2: $\Delta CCI_t = \alpha + \beta_1 \Delta CCI_{t-1} + \beta_2 \Delta CCI_{t-2}$

| | Res.Df | Df | F | Pr(>F) | |
|---|---|---|---|---|---|
| 1 | 123 | | | | |
| 2 | 125 | -2 | 2.4609 | 0.08955 | . |

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Therefore, the null hypothesis of the Granger test can be rejected that the lags of the Supervised Sentiment Index have no predictive power for the Consumer Confidence Index. This means that the Sentiment Index contains useful information about the future behavior of the Consumer Confidence Index, which confirms the hypothesis of the research.

### 3.4.3. Cointegration

As a final step the cointegration of the Sentiment Index[3] and the CCI is tested. The time series are cointegrated if a linear combination of them is stationary (Brooks, 2018). Cointegration suggests that the two-time series have a common long-term trend: they move together in the long run but short deviations occur. The cointegration was tested with the Engle-Granger Augmented Dickey-Fuller test for cointegration (based on Hanck et al., 2019). The test rejects the null hypothesis of nonstationarity at 1% level of significance for the difference of the time series, which suggests that the SI and the CCI are cointegrated. Future research could apply the Vector Error Correction Model (VECM) for the time series and utilize the cointegration of the time series.

**Table 11. Cointegration test for the SSI and CCI**

| **Augmented Dickey-Fuller Test Unit Root / Cointegration Test:** |
| --- |
| The value of the test statistic is: -3.9653 |



[3] Only the Supervised Sentiment Index is tested, since that was the better performing Index.

## 3.5. Discussion

The results of the ADL(2,2) model and the Granger Causality test showed that the Supervised Sentiment Index is a useful predictor of the Consumer Confidence Index. In addition, it was showed that the Supervised Sentiment Index is cointegrated with the CCI, which confirms that the series have a common long-term trend. These results confirm the hypothesis of the paper and shows that news sentiments contain information about the future behavior of consumer expectations. The findings of the paper are in line with the consensus in the current literature, but the published papers have more robust results and showed a stronger relationship.

On the other hand, the Dictionary-based Sentiment Index was not a useful predictor. It was expected that the results are sensitive to the method of the sentiment analysis, since the different methods could yield substantially different results. The dictionary used in this paper provides a raw estimate of the real sentiment of news, different sentiment dictionaries could improve this result. Regarding the supervised machine learning algorithm, a bigger training set could provide more precise results as well, and the coding of the training set could be verified by several human coders to ensure its quality. The research question also justifies a more in-depth analysis of the news in Hungary, where more than 4 news outlets could be examined.

The results are contributing to the news sentiment literature by conducting the analysis on Hungarian corpus and confirms that the findings that were shown on English texts are present in the Hungarian context as well. The results also suggest that conducting similar analyses could be relevant in other countries as well, since the relationship can be examined with limited resources as well.

Further research could use the findings of this paper to forecast the Consumer Confidence Index and evaluate the performance of the forecast. The method could be refined by including macroeconomic variables in the equation, such as lagged values of unemployment or inflation. A

related research question could focus on the prediction of macroeconomic variables in Hungary with the help of news sentiments, as it was examined in other languages but not in Hungarian yet.

# 4. Policy Implications

Nowcasting of certain variables is a challenge for policymakers, as several indicators' data is calculated and published with a lag. Having this information is often essential to have a clear view about the state of the economy, which is the foundation of sound economic policies.

Consumer expectations are also useful for policymakers since these influence purchasing and saving decisions. If policymakers misjudge the consumers' sentiments, then the enacted economic policies might not bring the desired results. Knowing if consumers are afraid of a recession or are expecting a boom will inform policymakers about whether policies that encourage spending, provide liquidity or encourage savings are required.

The results of the paper suggest that using sentiment analysis of economic news articles might serve as a tool to nowcast consumer sentiment. With the help of the Sentiment Index the predictions and forecasts about consumer expectations can be improved, which gives a more precise picture about future trends in a timely manner. In theory, if a team of data scientists monitored the tone and sentiment of economic news daily, they could estimate the consumers' sentiment each day. The resulting time series would be very noisy, a smoothed version of the raw daily series would be most useful.

The result of the paper also suggests that media coverage of events is crucial for policymakers, and a very important channel to mediate messages, where not only the substance of the policy matters but rather its appearance and perception in the news. Policymakers should monitor in the media the coverage of certain events and make conclusions about how the results were interpreted by the news outlets. The framing and journalists' and professionals' reactions matter for the consumers, and if the reactions are not in line with the policymakers' agenda, then the public communication could be fine-tuned.

# 5. Conclusion

The research set out to uncover whether economic news' sentiment is a helpful predictor of consumers' expectations and sentiment. Policymakers try to predict consumers' behaviour, for which their current expectations and sentiment provides a useful input. The current research provides an alternative method to the currently used surveys that measure consumer expectations.

The hypothesis of the research was tested by an Autoregressive Distributed Lag model with two lags; and Granger Causality. I tested the hypothesis both with the Dictionary-based Sentiment Index and the Supervised Sentiment Index. The results showed that the second lagged coefficient of the SSI was significant at 10% level in the ADL (2,2) model. When the model was tested with the DSI, the coefficients of the DSI were not significant. The SSI performed well in the Granger Causality estimate, the test showed that including the SSI improves the predictions of the CCI, and the result was significant at the 10% level. The Granger Causality estimate for the DSI was not significant. A cointegration test was conducted for the better performing SI (the SSI), which confirmed that the Supervised Sentiment Index and the Consumer Confidence Index follow a common long-term trend.

The ADL model and Granger Causality both suggested that the supervised machine learning-based Sentiment Index is a useful predictor of the Consumer Confidence Index. The hypothesis of the research was confirmed, which suggests that news sentiment is helpful in predicting consumer sentiment in Hungary. To the knowledge of the author this is the first paper to confirm this hypothesis on Hungarian text data, the paper is contributing to literature by providing results in a specific geographic context. The results also suggest that news sentiments are informative even if a few selected and relatively small news outlets are investigated, which could encourage the research of this question in several other countries.

Further research could fine-tune the sentiment analysis methods and improve these results. Including more news outlets and a longer time horizon or creating more sophisticated keyword

lists and training sets could uncover more about the research question. The predictive power of the Sentiment Index could also be tested by creating forecasts and evaluating the results.

The findings suggest that the framing and opinions written about certain events and policy changes bear importance, and policymakers could make use of the vastly available text data to improve forecasts. The more sophisticated text analysis tools make the processing of text data quicker, and with more Hungarian text mining tools nowcasting of Consumer Sentiment could be a reality in the near future.

# 6. References


Bram, J., Ludvigson, S.C., 1998. "Does consumer confidence forecast household expenditure? A sentiment index horse race." *Econ. Policy Rev.* 4 (2).

Brooks, C. (2008). *Introductory econometrics for finance*. Cambridge [England], Cambridge University Press.

Buckman, S.R., Shapiro, A.H., Sudhof, M. and Wilson, D.J., 2020. "News sentiment in the time of COVID-19". *FRBSF Economic Letter*, 8, pp.1-05.

Carroll, C.D., Fuhrer, J.C., Wilcox, D.W., 1994. "Does consumer sentiment forecast household spending? If so, why?" *Am. Econ. Rev.* 84 (5), 1397–1408.

Digitális Közönségmérési Tanács (2021): *Toplisták*, URL: https://dkt.hu, (Accessed on 24 April 2021)

Doms, M.E. and Morin, N.J., 2004. "Consumer sentiment, the economy, and the news media." *FRB of San Francisco Working Paper,* (2004-09).

Engelberg, J. E., & Parsons, C. A. 2011. "The causal impact of media in financial markets." *The Journal of Finance*, 66(1), 67–97.

Fang, L., & Peress, J. 2009. "Media coverage and the cross-section of stock returns." *The Journal of Finance*, 64(5), 2023–2052. http://dx.doi.org/10.1111/j.1540-6261.2009.01493.x.

Fraiberger, S., 2016. "News sentiment and cross-country fluctuations." *In Proceedings of the First Workshop on NLP and Computational Social* Science (pp. 125-131).

Garcia, D., 2013. "Sentiment during recessions." *The Journal of Finance*, 68(3), pp.1267-1300. http://dx.doi.org/10.1111/jofi.12027.

Garz, M. (2013). "Unemployment expectations, excessive pessimism, and news coverage." *Journal of Economic Psychology*, 34, 156–168. http://dx.doi.org/10.1016/j.joep.2012.09.007.

Griffin, J.M., Hirschey, N.H. and Kelly, P.J., 2011. "How important is the financial media in global markets?". *The Review of Financial Studies,* 24(12), pp.3941-3992.

Hanck, C., Arnold, M., Gerber, A. and Schmelzer, M., 2019. *Introduction to Econometrics with R. Essen*: University of Duisburg-Essen.

Hangya, V. and Farkas, R., 2017. "A comparative empirical study on social media sentiment analysis over various genres and languages". *Artificial Intelligence Review*, 47(4), pp.485-505.

Hester, J.B. and Gibson, R., 2003. "The economy and second-level agenda setting: A time-series analysis of economic news and public opinion about the economy". *Journalism & Mass Communication Quarterly*, 80(1), pp.73-90.

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.

Molnár, S., Molnár, M., Naár-Tóth, Z. and Timár, T., 2015. "Forecasting share price movements using news sentiment analysis in a multinational environment." *Hungarian agricultural engineering*, (28), pp.53-55.

Müller, A.C. and Guido, S., 2016. *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc.

Larsen, V. H., Thorsrud, L. A., 2019. "Business cycle narratives". *CESifo Working Paper Series* 7468, CESifo Group Munich.

OECD, 2021. Consumer confidence index (CCI) (indicator). doi: 10.1787/46434d78-en (Accessed on 27 May 2021)

Ostapenko, N., 2020. Macroeconomic expectations: news sentiment analysis (No. wp2020-5). Bank of Estonia.

Rambaccussing, D. and Kwiatkowski, A., 2020. "Forecasting with news sentiment: Evidence with UK newspapers." *International Journal of Forecasting*, 36(4), pp.1501-1516.

Sebők, M., Ring, O., Máté, Á., 2021. *Szövegbányászat és Mesterséges Intelligencia R-ben.* Budapest*: Typotext.*

Shapiro, A.H., Sudhof, M. and Wilson, D.J., 2020. "Measuring news sentiment." *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.07.053.

Sharpe, S., Sinha, N., & Hollrah, C., 2017. "What's the story? A new perspective on the value of economic forecasts". Finance and Economics Discussion Series 2017-107, Board of Governors of the Federal Reserve System (U.S.), revised 20 Aug 2018.

Soroka, S.N., 2006. "Good news and bad news: Asymmetric responses to economic information." *The journal of Politics*, 68(2), pp.372-385.

Souleles, N.S., 2004. "Expectations, heterogeneous forecast errors, and consumption: Micro evidence from the Michigan consumer sentiment surveys." *Journal of Money, Credit and Banking*, pp.39-72.

Starr, M., 2012. "Consumption, sentiment, and economic news." *Economic Inquiry*, 50(4), 1097–1111. http://dx.doi.org/10.1111/j.1465-7295.2010.00346.x.

Tetlock, P.C., 2007. "Giving content to investor sentiment: The role of media in the stock market." The Journal of Finance, 62(3), pp.1139-1168.

Tóth, J., 2020. "Negative and Engaged: Sentiments towards the 2016 Migrant Quota Referendum in Hungarian Online Media." East European Politics and Societies, 35(02), pp.493-518.

Uhl, M.W., 2010. "Explaining US consumer behavior and expectations with news sentiment" (No. 263). KOF Working Papers.

Vincze, V., 2014. "Uncertainty detection in Hungarian texts." In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1844-1853).

Vincze, V., 2016. "Detecting uncertainty cues in Hungarian social media texts." In Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM) (pp. 11-21).

# Appendix A.: Key-word list

The following key-word list was used to filter the dataset. Some of the words are stemmed and some words are included with different suffixes, which is a common method to handle the diversity of text data.

| **Keywords in Hungarian** | **Keywords in English (translated)** |
|---|---|
| ['GKI', | ['GKI', |
| 'Gazdaságvédelmi Alap', | Economic Protection Fund' |
| 'KSH', | 'KSH', |
| 'MNB', | 'MNB', |
| 'adókedvezmény', | tax credit' |
| 'adóteher', | 'tax burden', |
| 'adóterhek', | 'tax burdens', |
| 'albérletár', | housing prices', |
| 'anyagi gond', | 'financial trouble', |
| 'babakötvény', | 'baby bond', |
| 'befektet', | 'invest', |
| 'beruház', | 'invest', |
| 'bevételkiesés', | 'loss of income', |
| 'brexit', | 'brexit', |
| 'béremel', | 'raise wages', |
| 'bértámogatás', | 'wage subsidy', |
| 'büdzsé', | 'budget', |
| 'bővít', | 'expand', |
| 'családoknál marad', | families keep', |
| 'családtámogatás', | 'family support', |
| 'csok', | 'csok', |
| 'diákhitel', | 'student loan', |
| 'export', | 'export', |
| 'felfüggeszt', | 'suspend', |
| 'fizetőeszköz', | 'currency', |
| 'foglalkoztat', | 'employs', |
| 'fogyasztóbarát', | 'consumer friendly', |
| 'forintpiac', | 'forint market', |
| 'gazdasági növeked', | 'economic growth', |
| 'gazdasági visszaesés', | 'economic downturn', |
| 'gazdaságélénkít', | 'economic recovery', |
| 'gazdálkod', | econom', |
| 'gyarapod', | 'prosper', |
| 'gyártó üzem', | 'manufacturing plant', |
| 'hitelkeret', | 'credit line', |
| 'import', | 'import', |
| 'inflatór', | 'inflator', |
| 'infláció', | 'inflation', |

| | |
|---|---|
| 'ipari park', | 'industrial park', |
| 'jegybank', | 'central bank', |
| 'kapacitás', | 'capacity', |
| 'kedvezményes hitel', | preferential loan', |
| 'kereslet', | 'demand', |
| 'kompenzál', | 'compensates', |
| 'korszerűsít', | modernize', |
| 'kínálat', | 'supply', |
| 'költségtérít', | 'reimburses', |
| 'kötvény', | 'bond', |
| 'lakásfelújít', | 'home renovation', |
| 'lakásállomány', | property stock', |
| 'leállás', | 'downtime', |
| 'leépítés', | 'downsizing', |
| 'megtakarít', | 'save', |
| 'minimálbér', | 'minimum wage', |
| 'moratórium', | 'moratorium', |
| 'munkahely', | 'workplace', |
| 'munkanélküli', | 'unemployed', |
| 'nemzetgazdaság', | 'national economy', |
| 'nyereség', | 'profit', |
| 'nyugdíj', | 'pension', |
| 'olaj', | 'oil', |
| 'otthonteremt', | home creation', |
| 'pályázat', | 'competition', |
| 'recesszió', | 'recession', |
| 'rezsiköltség', | 'overheads', |
| 'segély', | 'aid', |
| 'spórol', | 'save', |
| 'szakszervezet', | 'trade union', |
| 'tao', | 'tao', |
| 'településfejlesztés', | 'urban development', |
| 'támogat', | 'support', |
| 'többletköltség', | 'additional cost', |
| 'tönkreme', | 'ruin', |
| 'törleszt', | 'repay', |
| 'tőkejövedel', | 'capital income', |
| 'versenyelőny', | 'competitive advantage', |
| 'veszteség', | 'loss', |
| 'vevő', | 'customer', |
| 'világgazdaság', | 'world economy', |
| 'visszaesés', | 'relapse', |
| 'visszatérít', | 'refund', |
| 'válság', | 'crisis', |
| 'válság', | 'crisis', |
| 'árfolyam'] | 'exchange rate'] |

# Appendix B: Examples for manual coding of articles

| Year | Title (Translated) | Manual coding | Comment to the manual coding | Source |
|---|---|---|---|---|
| **2010** | The exchange rate of the forint hardly changed | 0 (neutral) | descriptive, no opinion or feeling mediated | 24hu |
| **2010** | Ten candidates for the title of most successful business leader in the crisis | 0 (neutral) | Not relevant | 24hu |
| **2014** | Next year's student loan interest rates have been set | 0 (neutral) | descriptive, no opinion or feeling mediated | 24hu |
| **2015** | The CSO reviews the calculation of the subsistence minimum | 0 (neutral) | descriptive, no opinion or feeling mediated | 24hu |
| **2016** | MNB issues commemorative coins, one worth 20,000 | 0 (neutral) | Not relevant | 24hu |
| **2013** | The deficit has grown again, the population is diligently repaying | 0 (neutral) | Contains both positive and negative opinions | portfolio |
| **2013** | Soros: Europe can still fall apart easily | 0 (neutral) | Contains both positive and negative opinions | index |
| **2020** | Tourism jobs supported with 15 billion | 1 (positive) | Positive news (well-received economic policy) | index |
| **2010** | According to the finance minister, there will be no recession this year | 1 (positive) | Positive news (no recession) | index |
| **2011** | Here are the most optimistic numbers! | 1 (positive) | Positive news (economic growth) | portfolio |
| **2013** | More people work, less unemployed | 1 (positive) | Positive news (decreased unemployment) | origo |
| **2011** | It is cheaper to repay the Hungarian public debt | 1 (positive) | Positive news (low public debt service) | 24hu |
| **2010** | The investment at Sonkád will create 2,500 new jobs | 1 (positive) | Positive news (new jobs created) | 24hu |
| **2010** | Finnish investors are interested in Hungary | 1 (positive) | Positive news (economy attracts investors) | 24hu |
| **2010** | Simor: Every household is affected by the loss of investor confidence | -1 (negative) | Negative news (decreased investor confidence) | 24hu |
| **2013** | We no longer have the tools to deal with the crisis | -1 (negative) | Negative news (problems of crisis management) | index |
| **2019** | There is still a long way to go for the global economy to recover | -1 (negative) | Negative news (slow economic growth) | origo |
| **2011** | S&P has put Hungary on a negative watch list! | -1 (negative) | Negative news (decreased investor confidence) | portfolio |
| **2013** | Even if the industry improves, exports do not | -1 (negative) | Negative news (macroeconomic data) | portfolio |
| **2016** | The MNB's GDP forecast for this year is cracking | -1 (negative) | Negative news (macroeconomic data) | portfolio |
| **2020** | Experts: A 13-month pension is like throwing a lifebelt to a drowning person on the beach | -1 (negative) | Negative news (inadequate policy) | portfolio |