ASEC 2022

Proceeding Paper

Evaluation of News Sentiment in Economic Activity Forecasting †

Mantas Lukauskas [1,*], Vaida Pilinkienė [2], Jurgita Bruneckienė [2], Alina Stundžienė [2], Andrius Grybauskas [2] and Tomas Ruzgas [1]

[1] Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, 44249 Kaunas, Lithuania; e-mail@e-mail.com
[2] School of Economics and Business, Kaunas University of Technology, 44249 Kaunas, Lithuania; e-mail@e-mail.com (V.P.); e-mail@e-mail.com (J.B.); e-mail@e-mail.com (A.S.); e-mail@e-mail.com (A.G.)
* Correspondence: mantas.lukauskas@ktu.lt
† Presented at the 3rd International Electronic Conference on Applied Sciences; Available online: https://asec2022.sciforum.net/.

Featured Application: The analysis of news sentiments presented in this article can be used for various research and practical applications. In this study, news sentiment analysis is used in order to perform more accurate forecasting of economic indicators. Based on the results obtained in the study, it can be seen that the negative sentiment index is strongly linked to economic indicators. Based on the value of the sentiment, for this reason, economic activity and individual indicators of economic activity can be predicted. It can be used in the activities of various decision-makers, banks, and state institutions.

Abstract: Natural language processing is a rapidly expanding field of artificial intelligence, the main goal of which is linguistics. This field allows various mathematical/computer science techniques to be applied to natural language processing. Sentiment analysis is one of the most common tasks solved based on natural language processing. The primary purpose of sentiment analysis is to determine the mood (happy, sad, angry, and others) or polarity (negative, neutral, positive) of the presented text. Based on the relevance of the application of natural language processing, this study aims to create a dataset of Lithuanian news and determine the sentiment of this news. Identified news sentiment is associated with different indicators of economic activity. More than 1 million articles (1,256,227) have been collected from the largest news portal. The articles were collected from the first month of 2006 to the 5th of 2022. More than 20,000 different LSTM, GRU, RNN models were built with different parameters and data sets (univariate, individual sentiments, all sentiments, clusters). Based on the obtained results, it can be observed that the inclusion of sentiments in clustering increases the accuracy of forecasting different economic activity indicators. The highest accuracy in all cases is obtained based on the best sentiment for individual time series.

Keywords: clustering; economic activity; natural language processing; NLP; transformers; BERT; forecasting; nowcasting; economic sentiment

MSC: 68T50; 91B84; 62H30

Citation: Lukauskas, M.; Pilinkien, V.; Bruneckien, J.; Stundžien, A.; Grybauskas, A.; Ruzgas, T. Evaluation of News Sentiment in Economic Activity Forecasting. *Eng. Proc.* **2022**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 1 December 2022

1. Introduction

Natural language processing is a rapidly expanding field of artificial intelligence, the main goal of which is linguistics. This field allows various mathematical/computer science methods to be applied to natural language processing. Due to the rapid development of the field, natural language processing can be applied to a wide range of tasks: automatic translation [1], automatic correction of grammatical errors in texts [2], improving search engines based on natural language processing [3], voice-over text and many other areas [4]. Also, sentiment analysis is one of the most common tasks that is solved based on natural language processing. The primary purpose of sentiment analysis is to determine a given text's mood (happy, sad, angry, etc.) or polarity (negative, neutral, positive). Such natural language processing allows for processing millions of texts and determining the sentiment of these texts [5]. Sentiment analysis is often used in corporate activities to analyze comments about goods, companies or other items. Due to the high speed of analysis and possible determination of sentiment, sentiment analysis can also be used in economic activities when various publicly available sources are evaluated: Twitter messages, Facebook messages, reports in the press, others. In order to determine the sentiment of economic news/texts, specific words that can describe the sentiment are quite often used, and the number of these words is determined [6]. The authors often analyze such words and combinations of words as economic recession, falling shares, war, political uncertainty and others. An index is created based on the number of these words [7]. Recent scientific research proves the relevance of this field of science and application as more and more attention is paid to sentiment analysis applied to natural language processing. One such study is Shapiro et al. (2022), which applied both word-based computing and machine learning techniques [8]; other authors carried out similar studies.

Based on the relevance of the application of natural language processing, this study aims to create a dataset of Lithuanian news and determine the sentiment of this news. The determined news sentiment is associated with different indicators of economic activity. It is important to note that the data sets in the studies conducted by other authors are relatively small, so this study will allow us to evaluate the impact of an extensive data set on sentiment analysis. It is expected that this work will allow evaluation of the impact of a negative sentiment of different news categories on economic activity. Based on this objective, this paper hypothesizes that increasing negative news sentiment is inversely related to economic activity indicators. This scientific work would allow assessing the influence of negative sentiment of Lithuanian news on economic indicators. It would contribute to the development of this science field, help make more accurate forecasts and form further guidelines in the field of research.

2. Materials and Methods

This chapter reviews the materials and methods. In the first stage of this research, data was collected from the most prominent Lithuanian news portal. In order to collect this data, the Python package Playwright was used, which allows the data collection process to be automated. Each observation/text has three main parts: the title, the lead and the full text of the article. This research uses article titles and leads. Also, information such as publication date and article category (such as business, health, sports, and others) is collected from the articles. More than 1 million articles (1,256,227) have been collected from the largest news portal. The articles were collected from the first month of 2006 to the 5th of 2022. Below is the number of articles collected from different periods. Indicators of economic activity were collected using the database of the Lithuanian Statistics Department. Three leading indicators were chosen: monthly, annual, and production index. These data were chosen due to their current frequency, i.e., data are available monthly.

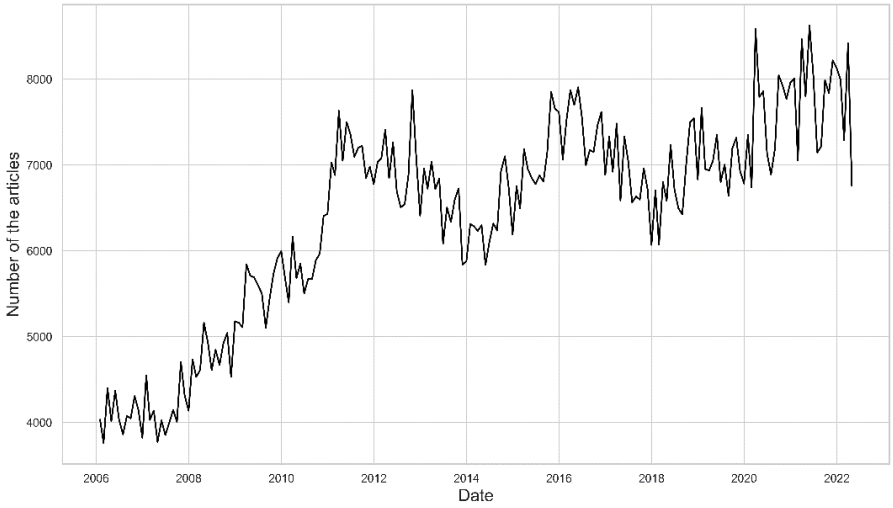

Figure 1. Number of articles over time.

This study uses pre-trained transform-er models for sentiment detection. In order to use the pre-trained transformer models, the collected texts were first translated into English. The translation into English was necessary because there are currently no existing quality models of the Lithuanian language. In order to ensure the accuracy of the translated texts, selected random texts were checked manually. Four different models of pre-trained transformers available in HuggingFace were used for sentiment analysis: DistilBERT-base-uncased, FinBERT, Twitter-roBERTa-base, and FinBERT-tone. The use of four different models is based on the fact that individual models may not always accurately determine sentiment due to the nature of their training, so combining different models allows for more accurate sentiment analysis and avoids individual models' influence. DistilBERT-base-uncased model is a reduced version of the Bert-base-uncased model but has exceptionally high performance. FinBERT is a model developed by Prosus, specifically designed to analyze financial texts. The third model used is FinBERT-tone, which is also trained on financial texts and determines financial texts' sentiment. [9]. This study's last but not most petite model is the Twitter-roBERTa-base model, specifically for sentiment analysis [10]. This model is trained using as many as 124 million Twitter messages collected over three years. The last model was chosen for its ability to identify sentiment in standard texts. The resulting overall sentiment index is calculated based on the results of all four models. The general sentiment index (SI) for time t is calculated according to the formula below:

$$SI_t = \frac{1}{N_t}\sum_{j=1}^{4}\sum_{i=1}^{N_t} T_j(A_{it}) \quad (1)$$

here SI_t is the sentiment index at a point in time t; T_j—a sentiment analysis model (transformer) is used since the sum of the four models used in total is up to 4, $T_j(A_{it})$ the output is given in the interval from 0 to 1; A_{it}—The ith news article at time t, where i is in the interval from 1 to N_t, where N_t is the number of news articles at a time t.

Below is the calculated sentiment index for the entire period under study. Based on the presented graph, it can be observed that the most significant increase in negative sentiment was observed after the start of the pandemic in early 2020 and also after the start of the war in Ukraine in early 2022.

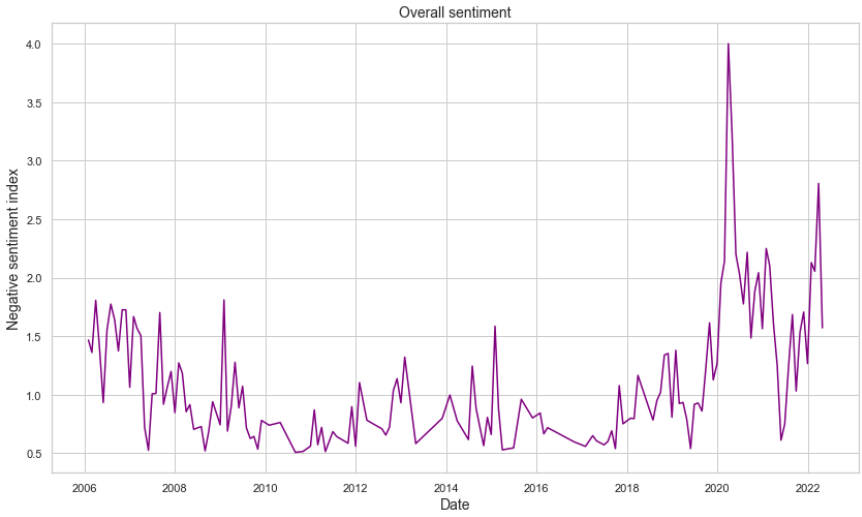

Figure 2. Negative overall sentiment over research period.

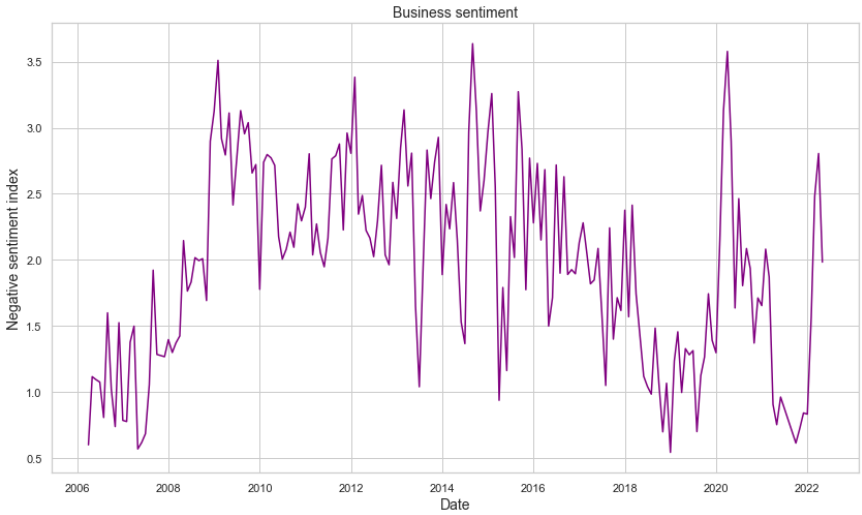

Figure 3. Negative business sentiment over the research period.

This study also uses different clustering methods to classify various texts into certain unknown groups. Clustering methods such as MIDE were used in this study [11], K-means, GMM and BGMM. Based on the clustering results, these results were also included in the forecasting of economic indicators.

3. Results

This section presents the main results of forecasting economic indicators based on sentiment and cluster analysis. These results are compared with the one-dimensional time series forecasting results. Economic indicators are forecasted using different lags (from 1 to 12 months). The following table presents the results of the best models for each economic activity indicator. The first row shows the results of univariate forecasting, the second row shows the forecasting based on the best sentiment, the third row shows the forecasting based on all sentiments, and the last row shows the forecasting based on cluster analysis. More than 20,000 LSTM, GRU, and RNN models were created with different parameters and datasets (univariate, individual sentiments, all sentiments, and clusters). Based on the obtained results, it can be observed that the inclusion of sentiments in clustering increases the accuracy of forecasting different economic activity indicators. The highest accuracy in all cases is obtained based on the best sentiment for individual time series. It is also interesting that although the largest cluster improves the forecasting

results, it is not better than the sentiment analysis results. This can be justified because only a tiny part of all texts belong to the largest cluster, so they do not necessarily reflect the exact situation.

Table 1. Results of economic activity forecasting based on univariate time series and multivariate time series.

Data/Model	RMSE		MAE		MAPE	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**
Monthly inflation rate						
Univariate	0.217	0.016	0.153	0.008	0.348	0.017
Overall sentiment	**0.134**	0.005	**0.097**	0.002	**0.243**	0.011
All sentiments	0.219	0.011	0.159	0.006	0.379	0.034
Biggest cluster	0.218	0.014	0.109	0.007	0.345	0.047
Annual inflation rate						
Univariate	0.097	0.015	0.064	0.015	0.303	0.050
Overall sentiment	**0.072**	0.009	**0.051**	0.007	**0.293**	0.045
All sentiments	0.116	0.061	0.088	0.048	0.301	0.127
Biggest cluster	0.126	0.067	0.087	0.051	0.305	0.168
Production index						
Univariate	0.224	0.013	0.187	0.015	0.446	0.041
Lithuania sentiment	**0.099**	0.005	**0.079**	0.005	**0.162**	0.012
All sentiments	0.152	0.012	0.088	0.011	0.177	0.013
Biggest cluster	0.166	0.024	0.093	0.022	0.198	0.022

4. Discussion

In this study, the main objectives were implemented. One of the primary and most difficult tasks was to collect an extensive data set and prepare it for further analysis. Although the collected data set was large enough, it is planned to use even more extensive data sets and collect articles from different news portals in the final study. After forecasting based on data clustering and sentiment analysis, it is noticeable that the results obtained in the case of clustering are not as good as when using sentiment analysis of all data. In the case of clustering, only the most significant data cluster was used, and for this reason, it can be assumed that the obtained results may not have been as accurate as if a more significant number of clusters had been used. Taking into account the results obtained during the research, it is observed that forecasting economic indicators based on sentiment analysis is better than forecasting one-dimensional data. Therefore, the hypothesis raised during the research is confirmed. This suggests that continuous monitoring of news and analysis of these news sentiments can allow for more accurate forecasts of economic indicators compared to one-dimensional data. As a further direction of this research, it is expected to train new sentiment analysis models explicitly adapted to the Lithuanian language, which would allow comparing the results with the results obtained in this study. It can also be assumed that using the entire text of the article instead of the article title and lead could positively impact the research results. For this reason, further studies with even larger datasets are planned.

Author Contributions: Conceptualization, M.L., V.P., J.B., A.S., A.G. and T.R; methodology, M.L., V.P., A.G. and T.R.; software, M.L. and T.R.; validation, V.P., J.B., A.S. and A.G.; formal analysis, M.L.; investigation, M.L., V.P. and A.G.; resources, M.L., V.P., J.B., A.S., A.G. and T.R.; data curation, M.L.; writing—original draft preparation, M.L. and A.G.; writing—review and editing, M.L., V.P., J.B., A.S., A.G. and T.R.; visualization, M.L.; supervision, V.P., J.B., A.S. and T.R.; project administration, V.P., J.B. and A.S.; funding acquisition, V.P., J.B. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from European Regional Development Fund (pro-ject No 13.1.1-LMT-K-718-05-0012) under a grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to the COVID-19 pandemic.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the area editor and the reviewers for giving valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y.; Zhang, J.; Zhai, F.; Xu, J.; Zong, C. Three strategies to improve one-to-many multilingual translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 31–4 November 2018; pp. 2955–2960.
2. Napoles, C.; Sakaguchi, K.; Post, M.; Tetreault, J. Ground truth for grammatical error correction metrics. In Proceedings of the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 588–593.
3. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2022**, 1–32.
4. Mittal, Y.; Toshniwal, P.; Sharma, S.; Singhal, D.; Gupta, R.; Mittal, V.K. A voice-controlled multi-functional smart home automation system. In Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 17–20 December 2015; pp. 1–6.
5. de Oliveira, N.R.; Pisa, P.S.; Lopez, M.A.; de Medeiros, D.S.V.; Mattos, D.M. Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information* **2021**, *12*, 38.
6. Taj, S.; Shaikh, B.B.; Meghji, A.F. Sentiment analysis of news articles: A lexicon based approach. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 30–31 January 2019; pp. 1–5.
7. Baker, S.R.; Bloom, N.; Davis, S.J. Measuring economic policy uncertainty. *Q. J. Econ.* **2016**, *131*, 1593–1636.
8. Shapiro, A.H.; Sudhof, M.; Wilson, D.J. Measuring news sentiment. *J. Econom.* **2020**, *228*, 221–243.
9. Huang, A.; Wang, H.; Yang, Y. FinBERT—A Deep Learning Approach to Extracting Textual Information. 2020. Available online: https://ssrn.com/abstract=3910214 (accessed on).
10. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. *arXiv* **2019**, arXiv:1912.00741.
11. Lukauskas, M.; Ruzgas, T. A New Clustering Method Based on the Inversion Formula. *Mathematics* **2022**, *10*, 2559.