Economics Letters 237 (2024) 111653

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Construction and analysis of uncertainty indices based on multilingual text representations

Viktoriia Naboka-Krell

Justus Liebig University Giessen, Licher street 64, Giessen, 35394, Hessen, Germany

ARTICLE INFO

Keywords:
Text-as-data
fastText embeddings
BERT
Economic policy uncertainty
Natural language processing

ABSTRACT

The work by Baker et al. (2016), who propose a dictionary based method and estimate the level of *economic policy uncertainty* (EPU) based on the occurrence of specific terms in ten leading newspapers in the USA, is among the first ones to detect the potential of text data in economic research. Following this line of research, this paper proposes automated approaches to construction of EPU indices for different countries based on newspapers' texts. Multilingual fastText word embeddings, (S)BERT embeddings, and a novel multilingual topic modeling approach are used to construct EPU indices for Germany, Russia, and Ukraine. It is shown that constructed EPU indices based on multilingual word embeddings are Granger causal to the economic activity in all of the considered countries.

1. Introduction

The work by Baker et al. (2016) is among the first ones to detect the potential of text data in economic research. Although dictionary based methods as in Baker et al. (2016) are widely used due to their simplicity and interpretability, recent advances in NLP offer many further possibilities to gain insights from text data. New approaches include the use of topic models such as Latent Dirichlet Allocation (LDA). Azqueta-Gavaldón (2017) proposes an LDA based procedure to build an uncertainty index that strongly resembles the index introduced by Baker et al. (2016) (BBD) index. The proposed EPU index is the aggregated time series based on the time series of the identified EPU related topics.

Some of them also make use of word embeddings, for example, to extend the EPU related term set as proposed by Ghirelli et al. (2019) for the case of Spain. The authors show that an unexpected shock in their modified EPU index leads to a significant decline in GDP, private consumption, and investments. Algaba et al. (2020) follow this approach and construct an EPU index for Belgium using GloVe word embeddings (Pennington et al., 2014). It has been shown that the constructed index negatively correlates (−0.62) with Consumer Confidence Indicator (CCI). Xie (2020) proposes a fully automated method to build an uncertainty index. The author applies the Wasserstein Index Generation model and uses word vectors to represent the analyzed text units in a vector space.

These examples show that word embeddings might have a considerable impact on future applications and methods in economic literature. In contrast to other methods, word embeddings are able to capture the semantic and syntactic characteristics of words, which is very useful in numerous cases. The current work is dedicated to examination of word and text representations in context of EPU measurement and contributes to the growing area of text-as-data applications in economics, particularly uncertainty measurement, in several ways. First, it proposes several approaches to construction of EPU indices in the multilingual setting without any supervision. Second, it applies a novel zero-shot topic modeling approach that allows to train a topic model in one language and to predict topic distributions for documents in unseen languages. Third, the resulting uncertainty indices are evaluated with regard to their impact on economic activity in selected countries.

2. Text representation techniques

Multilingual word embeddings are word vectors in multiple languages that are embedded in a shared vector space. These representations are characterized by the interpretability of the distances between them in different languages, meaning that similar words are closer to each other in the shared vector space. Several approaches have been proposed to train such multilingual word embeddings. One of the widely used approaches is the mapping based approach that relies on so-called off-the-shelf lexicons. Freely available multilingual fastText[1] word representations were also learned following the mapping based approach proposed by Joulin et al. (2018) (bilingual mapping) and Grave et al. (2018) (multilingual mapping by defining a pivot language).

E-mail address: Viktoriia.Naboka@wirtschaft.uni-giessen.de.

[1] fastText is a free library for text classification and representation learning.

https://doi.org/10.1016/j.econlet.2024.111653
Received 15 December 2023; Received in revised form 8 March 2024; Accepted 8 March 2024
Available online 11 March 2024

A great breakthrough in and a major contribution to the field of language model learning has been made with the publication of the work by Devlin et al. (2019). The authors present their novel approach to text representations BERT which differs substantially from existing models. BERT stands for Bidirectional Encoder Representations from Transformers and consists of a multi-layer Transformer encoder. BERT became the state-of-the-art in many NLP tasks. However, to overcome some capacity and time issues, Sentence BERT (SBERT) was introduced that was fine-tuned for semantic similarity search (Reimers and Gurevych, 2019).

Probabilistic topic modeling approaches are one further well-known and widely used tool for extracting and analyzing latent themes behind the underlying unstructured text data in different areas. Bianchi et al. (2021) introduce a novel approach to topic modeling for the multilingual setting. Multilingual contextualized topic modeling allows to train a topic model in one model and to infer topic distributions for documents in unseen languages just relying on their SBERT representations.

Further details on methods used in this paper are provided in Appendix A.

3. Data

For the empirical analysis, three datasets of news articles in three different languages are used: DER SPIEGEL for Germany, Lenta.ru for Russia, and UNIAN for Ukraine. The following preprocessing steps were taken: removing punctuation, numbers, special characters, stopwords and lowercase. The final datasets contain 833,454 articles in the period from January 2000 to September 2020 for Germany, 864,481 articles in the period from September 1999 to September 2020 for Russia, and 785,750 articles in the period from January 2007 to September 2020 for Ukraine. Economic activity is measured by industrial production index, which is often used in academic literature as a high-frequency indicator of a country's economic activity.

4. Construction of uncertainty indices

Overall, four different approaches are proposed to identify articles related to uncertainty in economic policy. The first approach is a dictionary based one, which uses fastText multilingual word embeddings to either identify three term sets referring to the three components of the EPU concept (later referred to as $dic1$) or one combined term set that should describe the EPU concept as a whole (later referred to as $dic2$). While the former method searches for nearest neighbors to the three terms *economic*, *policy*, and *uncertainty*, the latter makes use of the additive feature of the word vectors and searches for nearest neighbors to the compound word vector *economic* + *policy* + *uncertainty*. The similarity of the word vectors is measured by cosine similarity.[2] The number of relevant words is controlled for automatically based on a threshold, namely the 99.99% percentile of all the cosine similarity values between a certain word in one language (e.g. English word "policy") and all the word vectors available in other language (e.g. German) as shown in Fig. 1. The identified EPU related terms are presented in Appendix A.

The second approach considers the articles as Bag-of-Words and represents them as the sum of the constituent words vectors, which results in aggregated document embeddings (later referred to as *art_emb* approach). The third approach applies Transformer based text embeddings to identify articles that relate to EPU (later referred to as *art_sbert*).[3] Finally, a novel language agnostic topic modeling technique called zero-shot cross-lingual topic modeling is applied. Thereby, a topic model was trained based on German SBERT embeddings of articles and the topic distributions for Russian and Ukrainian articles were then also inferred based on SBERT embeddings of articles. EPU related topics have been identified using the embeddings of the most frequent topic words. In the following, this approach is referred to as $MCTM\_\{k\}\_Topic$. k can stand for the topic number/label of a topic that is identified as an EPU related topic or have the designation *combined*, if the average topic frequency of all the EPU related topics is used. Overall, 10 different EPU time series are provided for each country.

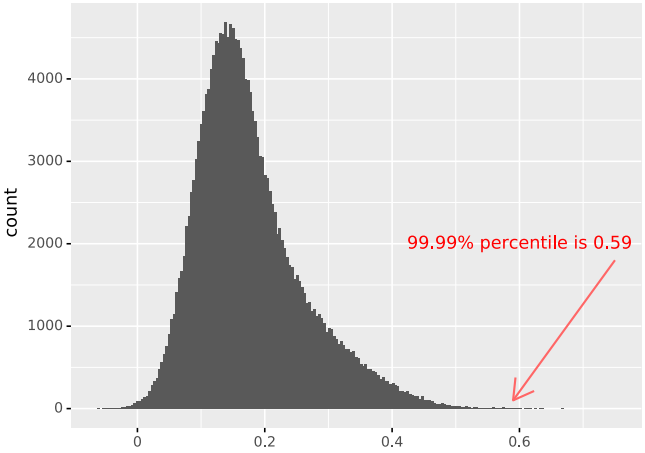

Fig. 1. Cosine Similarity Values between *policy* and all German words.

5. Results

5.1. EPU indices

This section presents the constructed indices. All the indices were normalized to have a mean of 100 and a variance of 1.

Fig. 2 shows the indices resulting from the $dic1$ and $dic2$ approaches. In Germany, the peaks correspond with such events as the September 11 attacks, economic crisis in Germany, global financial crisis, and Corona virus outbreak. The spikes of the EPU in Russia between 2004 and 2005 as well as in the period from 2014 and 2018 could be explained by the Orange Revolution in neighboring Ukraine and the Russia–Ukraine gas disputes, and by the Crimean crisis and the War in Donbass, respectively. Surprisingly, both of the constructed EPU indices for Ukraine show a downward trend. Some peaks can be identified at the beginning of 2007 (political crisis in Ukraine), in 2008-09 (global financial crisis), 2014 (beginning of the Crimean crisis and the War in Donbass), and at the beginning of 2019 (presidential and parliamentary elections). The Corona virus outbreak, instead, seems to have caused a relatively small increase in the uncertainty index. As this approach largely relates to that proposed by Baker et al. (2016), further analyses between the constructed indices and the available indices by Baker et al. (2016) (BBD indices) have been carried out (see Appendix A).

Figs. 3 and 4 show the *art_emd* and *art_sbert* indices for all three countries, respectively. There are some noticeable differences between

[2] Cosine similarity is defined as the cosine of the angle between two vectors. The values range from 0 to 1. A cosine similarity value of 1 means that the vectors are pointing in the same direction.

[3] To train SBERT articles' embeddings, a pre-trained `distiluse-base-multilingual-cased-v2` model was used. Thereby, Python's implementation of Sentence BERT (SBERT), namely *Sentence-Transformers*, was used to load and apply the model.

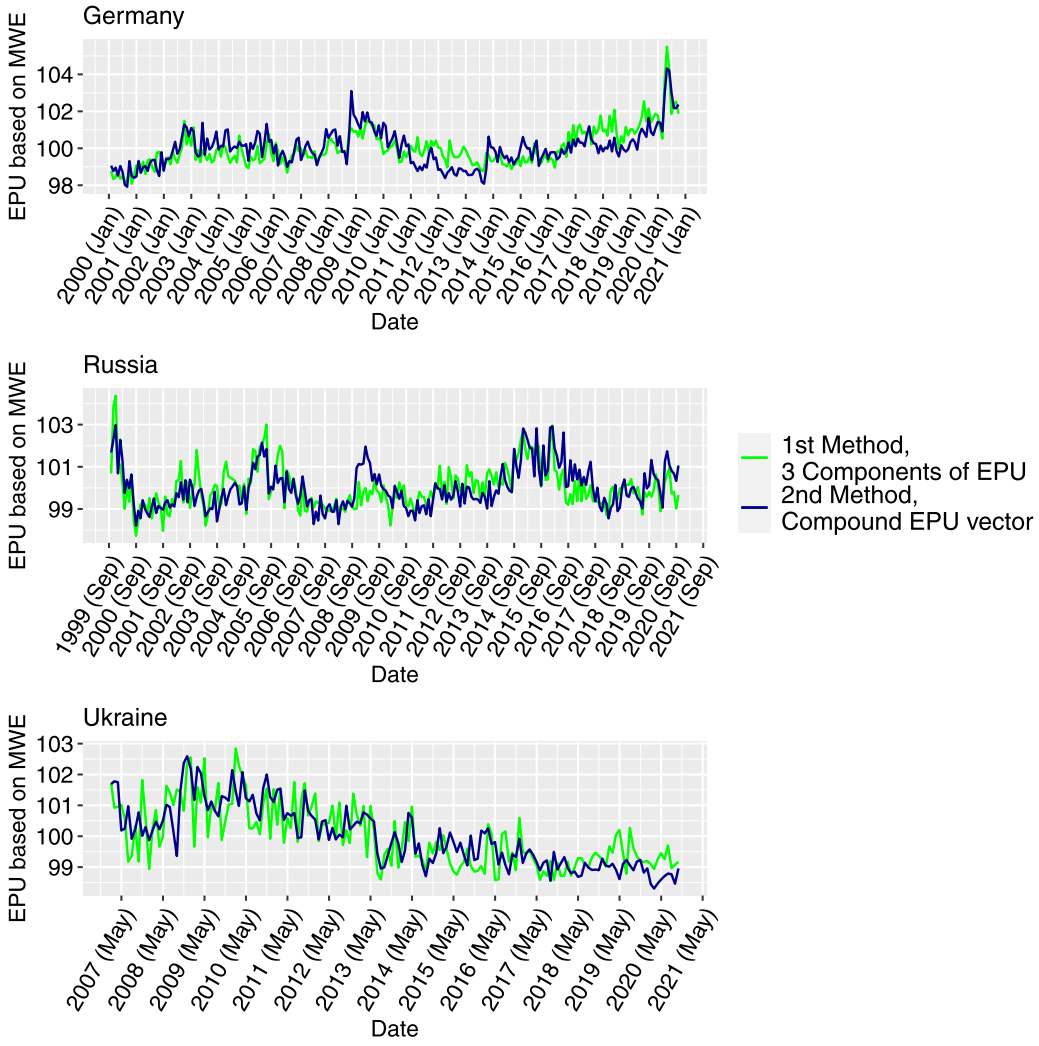

Fig. 2. *dic*1 and *dic*2 EPU Indices.

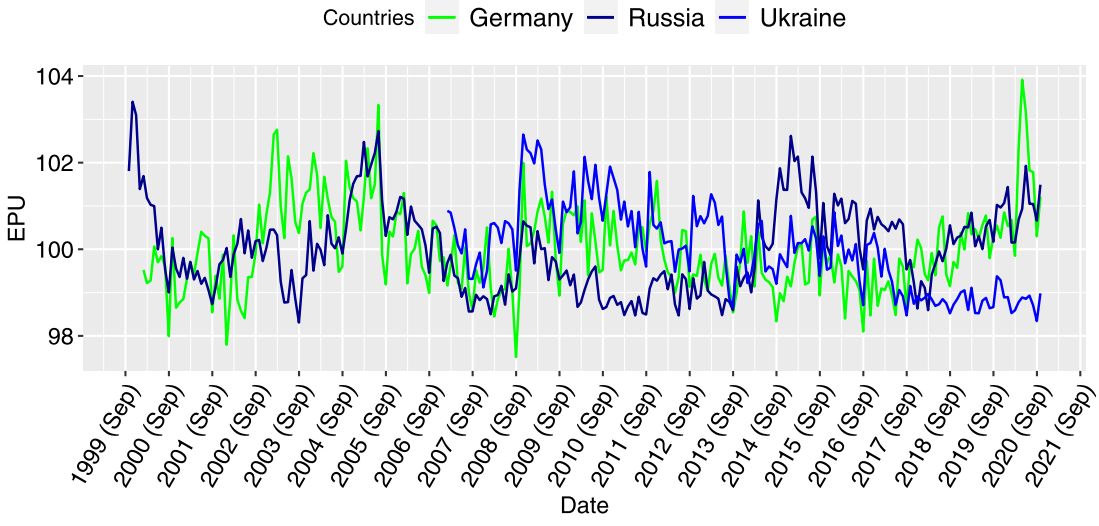

Fig. 3. *art_emd* EPU Indices.

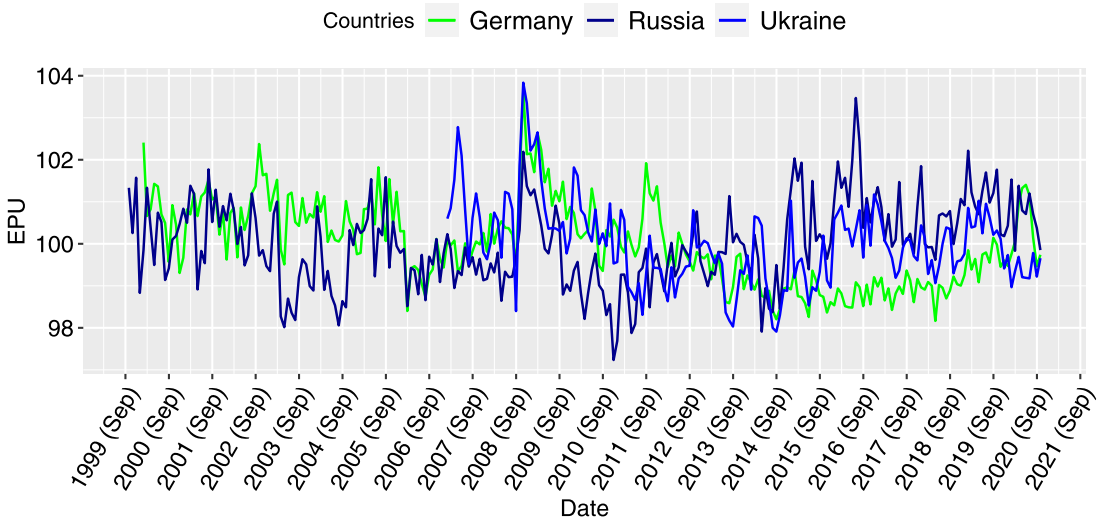

Fig. 4. *art_sbert* EPU Indices.

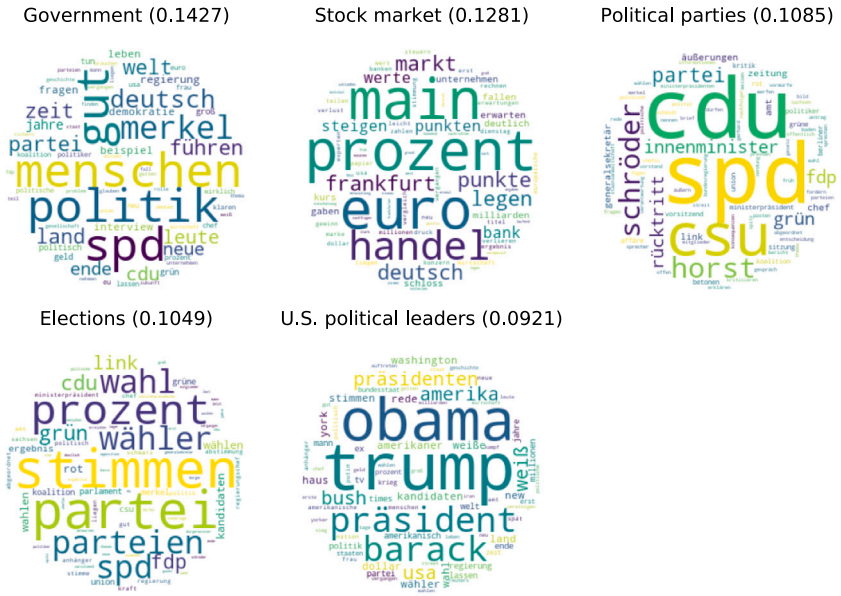

Fig. 5. MCTM with 40 Topics: EPU Related Topics.

the two, as for example, stronger interdependencies between Russia and Ukraine according to the *art_sbert* approach.

Finally, based on the results of the multilingual topic modeling five EPU related topics are identified in the current application. These are presented in Fig. 5. Based on the qualitative assessment of the most common words of the topics, these were assigned the following labels: `government`, `stock market`, `political parties`, `elections`, `U.S. political leaders`. The values in brackets represent the cosine similarity values to the EPU embedding. The corresponding time series for each country as well as additional robustness checks using U.S. data are presented in Appendix A.

5.2. VAR models

All the constructed EPU indices are tested within VAR models with regard to their impact on the economic activity of the countries. The economic activity is measured by the industrial production index, which is often used in academic literature as a high-frequency indicator of a country's economic activity (Baker et al., 2016; Perić and Sorić, 2018; Čižmešija et al., 2017).[4]

For each country, 10 two-dimensional VAR models with seasonal dummies were estimated, each including one of the constructed EPU indices and the corresponding industrial production index.[5]

Granger causality

The first set of analysis is dedicated to Granger causality tests, especially the null hypothesis "EPU does not Granger cause Industrial Production Index". According to Granger causality tests, the following EPU indices led to significant results at least in one country: $dic1$ (Ukraine), $dic2$ (all countries), *art_emb* (Germany, Ukraine), *art_sbert*

[4] The data for the analyses come from State Statistics Service of the considered countries.

[5] According to the performed stationarity tests, all the variables needed to be transformed to become stationary. For this reason, the first log differences of all the variables were calculated and used in all estimated Vector Autoregressive (VAR) models.

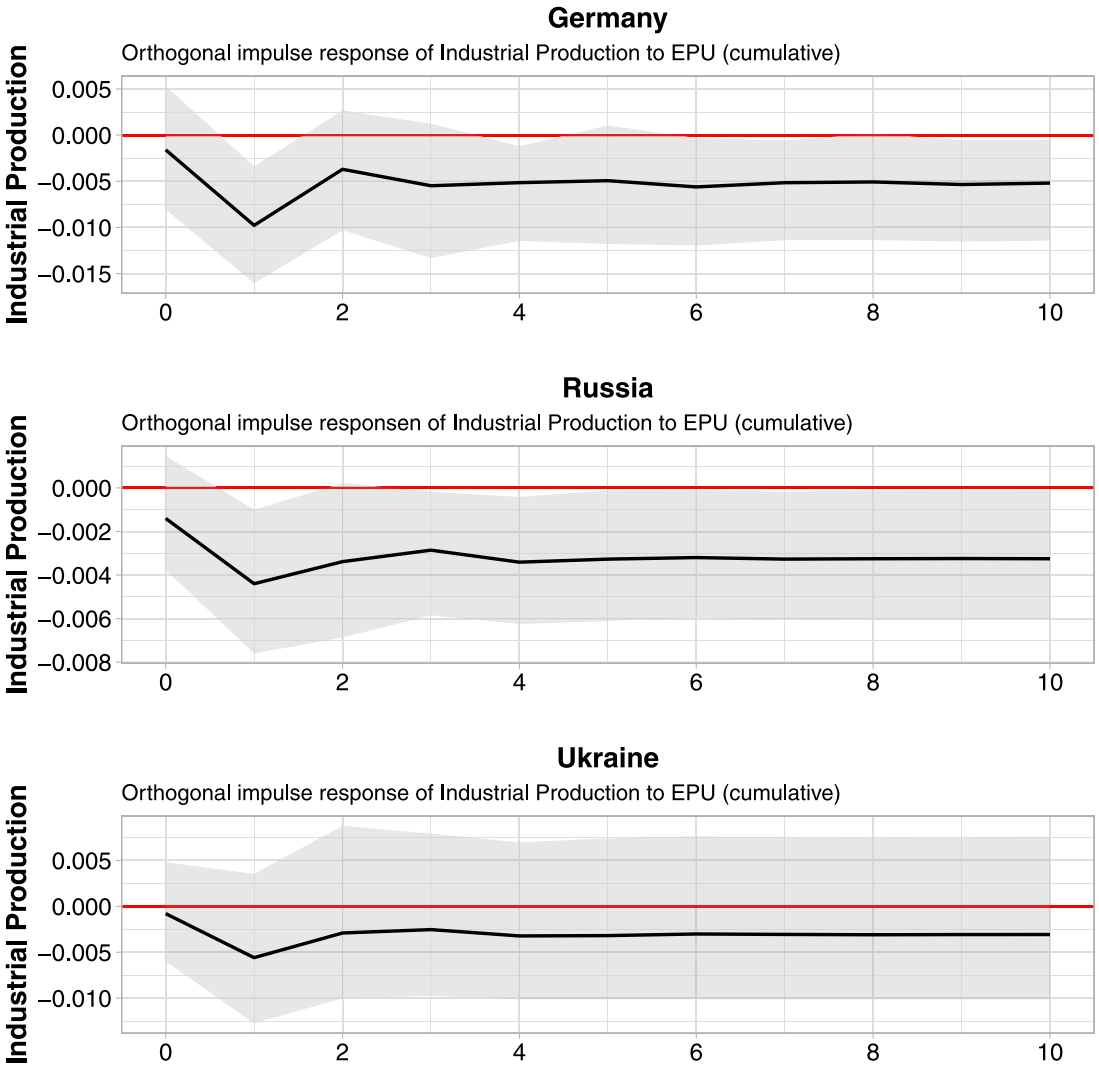

Fig. 6. IRFs: $dic2$ EPU Index →Industrial Production Index.

(Germany), $MCTM\_stock\_market\_Topic$ (all countries).[6] The results of Granger causality test are summarized in Appendix A.

Impulse response functions

A close look is taken on the $dic2$ EPU index, as this index has proved to be Granger causal to economic activity in all the considered countries. Fig. 6 illustrates the responses of the industrial production indices to an $dic2$ EPU indices shock in all considered countries. The shaded areas represent the 95% confidence bands. Thereby, the orthogonal impulse responses are considered meaning that contemporaneous effects are allowed. It can be inferred from the figure that one standard deviation shock in EPU leads to a significant drop of 0.1 and 0.44 percentage points in the industrial production index after one month in Germany and Russia, respectively. While in Germany there is only a short-term impact of the EPU shock on the industrial production, there is also a long-term significant negative impact of the EPU shock on the industrial production index (about 0.3 percentage points) in Russia. The pattern of the Impulse Response Function (IRF) in Ukraine is similar but not significant over the entire period.

6. Conclusions

One of the key finding of the current work is that the dictionary based approach in combination with multilingual word embeddings results in indices that are Granger causal to the economic activity in all of the considered countries. Further, to the best of my knowledge, the current paper is the first to apply a novel language agnostic topic modeling technique introduced by Bianchi et al. (2021) in economic context. One of the identified EPU related topics has proved to Granger cause the economic activity in all of the considered countries. This is one promising finding as the topic modeling approach by Bianchi et al. (2021) allows to predict topic distributions for texts in unseen languages just based on SBERT embeddings without renewed training of the model. Finally, it has been also shown that a sudden shock in the constructed EPU indices leads to significant short-term and/or long-term declines in industrial production. This finding indicates that the constructed EPU indices could be used as high frequency indicators of economic activity.

Data availability

Data will be made available on request.

[6] Results were considered significant if p-value is smaller than 10%.

Acknowledgments

I thank my supervisor Peter Winker and my colleagues for their support and advice.

Funding

No funds, grants, or other support was received.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2024.111653.

References

Algaba, A., Borms, S., Boudt, K., van Pelt, J., 2020. The economic policy uncertainty index for Flanders, Wallonia and Belgium. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.3580000, BFW digitaal/RBF numérique 2020/6, URL: https://ssrn.com/abstract=3580000.

Azqueta-Gavaldón, A., 2017. Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. Econom. Lett. 158, 47–50. http://dx.doi.org/10.1016/j.econlet.2017.06.032, https://www.sciencedirect.com/science/article/pii/S0165176517302598.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring Economic Policy Uncertainty*. Q. J. Econ. 131 (4), 1593–1636. http://dx.doi.org/10.1093/qje/qjw024.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E., 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. CoRR abs/2004.07737, arXiv:2004.07737.

Čižmešija, M., Lolić, I., Sorić, P., 2017. Economic policy uncertainty index and economic activity: What causes what? Croatian Oper. Res. Rev. 8 (2), 563–575.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://www.aclweb.org/anthology/N19-1423.

Ghirelli, C., Pérez, J.J., Urtasun, A., 2019. A new economic policy uncertainty index for Spain. Econom. Lett. 182, 64–67, https://ideas.repec.org/p/bde/wpaper/1906.html.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T., 2018. Learning Word Vectors for 157 Languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, URL: https://www.aclweb.org/anthology/L18-1550.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E., 2018. Loss in translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 2979–2984. http://dx.doi.org/10.18653/v1/D18-1330, URL: https://www.aclweb.org/anthology/D18-1330.

Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global Vectors for Word Representation. In: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. http://dx.doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162.

Perić, B.Š., Sorić, P., 2018. A note on the "Economic Policy Uncertainty Index". Soc. Indic. Res. 137 (2), 505–526.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, pp. 3982–3992. http://dx.doi.org/10.18653/v1/D19-1410, URL: https://www.aclweb.org/anthology/D19-1410.

Xie, F., 2020. Wasserstein index generation model: Automatic generation of time-series index with application to economic policy uncertainty. Econom. Lett. 186, 108874. http://dx.doi.org/10.1016/j.econlet.2019.108874, https://www.sciencedirect.com/science/article/pii/S0165176519304410.