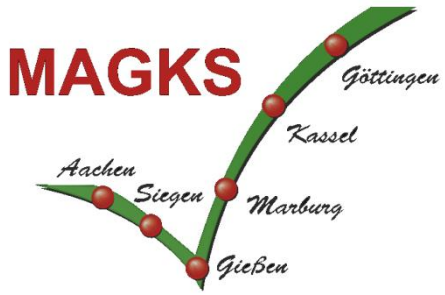

Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 10-2023

Viktoriia Naboka-Krell

Construction and Analysis of Uncertainty Indices based on Multilingual Text Representations

This paper can be downloaded from

https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Construction and Analysis of Uncertainty Indices based on Multilingual Text Representations

Viktoriia Naboka-Krell

Department, Justus Liebig University Giessen, Licher Str. 64, Giessen, 35394, Hessen, Germany.

Contributing authors: viktoriia.naboka@wirtschaft.uni-giessen.de;

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. First, multilingual fastText word embeddings and BERT text embeddings are used in order to define relevant EPU key words and EPU related articles, respectively. Further, multilingual conceptualized topic modeling introduced by Bianchi et al. (2021) is performed and EPU related topics are detected. It is shown that the constructed EPU indices based on fastText embeddings Granger cause the economic activity in all of the considered countries, namely Germany, Russia, and Ukraine. Also, some of the topics uncovered by multilingual conceptualized topic modeling have proved to Granger cause the economic activity in all of the considered countries.

Keywords: text-as-data, fastText emeddings, BERT, economic policy uncertainty, natural language processing

1 Introduction

Past and recent developments and events in the world such as the Great Recession in 2008, the Arab Spring in 2010, and the most recent Covid-19 pandemic have highlighted a considerable role of uncertainty in an economy and its impact on individuals, businesses, and economies. With a growing amount of text data available, the interest for construction text based uncertainty indices is increasing constantly.

A seminal work in this field is the contribution by Baker et al. (2016). The authors propose a dictionary-based method in order to construct an Economic Policy Uncertainty (EPU) index from ten leading news papers in the USA. Thereby, they define specific terms describing the EPU concept, namely *economy*, *policy*, *uncertainty* and count the number of occurrences of these specific terms in the defined period of time. Other researchers follow and/or refine their approach (Ghirelli et al., 2019; Algaba et al., 2020) as well as propose completely new ones (Azqueta-Gavaldón, 2017; Manela and Moreira, 2017; Xie, 2020). New approaches include the use of topic models such as Latent Dirichlet Allocation (LDA) (Azqueta-Gavaldón, 2017) and regression based approaches such as Support Vector Regression (SVR) (Manela and Moreira, 2017). The most recent work among the machine learning application is the one by Lolić et al. (2022) who propose ensemble learning approaches to improve the accuracy of the EPU index. The authors show that ensemble learning based approach outperforms the standard EPU indices as measured by correlation with standard uncertainty proxies.

In the Natural Language Processing (NLP) field, word and text vectors, also called embeddings, have been attracting increasing attention. Recently, such word embeddings have been also used to measure the diffusion of innovations (Lenz and Winker, 2020) and climate change transition risk (Kapfhammer et al., 2020). In the context of EPU measurement, some researchers also make use of word embeddings, for example, to extend the EPU related term set (Ghirelli et al., 2019) or to represent the analysed text units in a vector space (Xie, 2020). A recent study by Kaveh-Yazdy and Zarifzadeh (2021) address the problem of keyword sensitivity and proposes word embedding-based method for selecting relevant EPU related keywords for Iran. Miranda-Belmonte et al. (2023) successfully apply word embeddings in combination with topic modeling to construct real-time EPU index for the U.S. and Mexico.

The current work contributes to the growing area of text-as-data applications in economics in at least three ways. First, it proposes several approaches for constructing EPU indices in a 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, using Vector Autoregressive (VAR) models, the resulting uncertainty indices are evaluated with regard to their impact on economic activity in selected countries and, thus, their ability to serve as high-frequency indicators of important low-frequency macroeconomic variables.

The remainder of this paper is structured as follows. In Section 2, I will give an overview on the methods for text representation used in the current work. Section 3 presents the text data used in the current work as well as the process of constructing EPU indices for three countries. Section 4 briefly describes how the constructed indices are evaluated within VAR models. Finally, Section 5 describes the resulting indices as well as shows the results of the VAR analysis. The last section summarizes the main findings.

2 Text Representation Techniques

This section briefly describes different text representation techniques frequently used in different applications. Subsection 2.1 introduces (multilingual) word embeddings. A novel text representation approach called BERT is presented in Subsection 2.2. Finally, recently introduced multilingual topic modeling approach is described in Subsection 2.3.

2.1 Multilingual Word Embeddings

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 semantically 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. Thereby, two or more languages are projected into a joint vector space. In this work, I use multilingual word embeddings trained following the approach presented by Conneau et al. (2017).[1] These word embeddings were obtained using fastText[2] monolingual word embeddings.

2.2 (S)BERT

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 representation called BERT which differs substantially from existing models. BERT stands for Bidirectional Encoder Representations from Transformers and consists of a multi-layer Transformer encoder. Transformer could be seen as a black box that originally had a task of translating a piece of text in one language into another language. Later, BERT was fine-tuned to perform different downstream tasks such as sentence classification or sentiment analysis and became the state-of-the-art in many NLP tasks. In 2019, Reimers and Gurevych introduced Sentence BERT (SBERT) that was fine-tuned for semantic similarity search. Focusing on this one specific task enabled to overcome some capacity

[1] The vectors can be downloaded from Github under https://github.com/facebookresearch/MUSE#multilingual-word-embeddings.

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

and time issues. Reimers and Gurevych (2020) also extended their model to the multilingual setting. The authors highlight that the proposed approach allows meaningful representations of sentences in several languages from various language families.

2.3 Multilingual Contextualized Topic Modeling

Probabilistic topic modeling approaches are well-known and widely used in different areas. The main goal is to extract and analyse latent themes behind the underlying unstructured text data (Blei, 2012). The corpus is usually represented as the so-called Bag-of-Words (BoW) and a document-term matrix is constructed by counting the occurrence frequency of each term in each document of the considered corpus. Since such an approach is suitable for the analysis of almost all monolingual corpora from different domains, applying it to a multilingual corpus would result in topics separated by language.

Development of new natural language representation techniques like (S)BERT presented in the previous subsection enable new approaches to topic modeling. Bianchi et al. (2021) introduce a novel approach to topic modeling for the multilingual setting, namely Multilingual Contextualized Topic Modeling (MCTM). MCTM allows to train a topic model in one language and to infer topic distributions for documents in unseen languages just relying on their SBERT vector representations. In quantitative and qualitative analyses, Bianchi et al. (2021) have shown that the proposed MCTM approach leads to coherent topics and allows meaningful assignment of the documents in unseen languages to trained topics in another language.

3 Data and Construction of Uncertainty Indices

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.

3.1 EPU Indices based on Multilingual Word Representations

The first approach follows the method described by Baker et al. (2016). However, instead of defining the EPU related terms manually, I propose to use pre-trained multilingual word vectors described in Subsection 2.1 to either identify three term sets referring to the three components of the EPU concept or one combined term set that should describe the EPU concept as a whole. While the former method searches for nearest neighbors to the three

terms *economic*, *policy*, and *uncertainty* in the shared multilingual word vector space, 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.[3] 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 Figure 1.[4] The identified EPU related terms are presented in Appendix A.

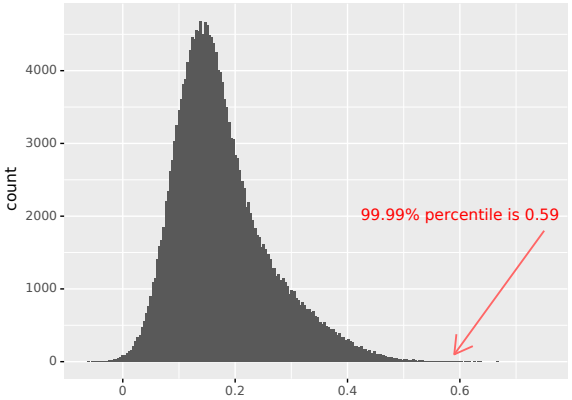

Figure 1: Cosine Similarity Values between *policy* and all German words

An article is then labelled as an EPU article if it contains at least one of the terms from each category for the first alternative (*economic*, *policy*, *uncertainty*) or at least one term from the compound EPU term set for the second alternative. Finally, the constructed EPU indices represent the monthly share of EPU articles.

3.2 EPU Index based on Aggregated Articles' Representations

The next approach involves the whole articles content by constructing articles embeddings. These articles emebeddings represent the sum of the constituent parts of an article, i. e. constituent word embeddings. Thereby, the BoW article representations are used. Even though the semantic and syntactic relations between the words in the BoW articles representations are neglected, some

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

[4]This choice is preferred to, for example, a *k nearest neighbours* approach, as this allows more flexibility. Russian and Ukrainian languages are characterized by strong inflexional structure. Thus, just taking five nearest neighbours to a certain word would often result in the words with the same word stem.

context information is provided, as the embeddings of all the words in an article are taken into account. Thus, the expectation here is that although texts might not contain specific words defined as EPU terms, the occurrence of certain words together in one document might result in an embedding close to the EPU embedding.

Then, cosine similarity values between all the aggregated articles and the sum vector *economic + policy + uncertainty* are calculated. The articles the cosine similarity of which is higher than or equal to the 90%[5] percentile of all the cosine values are labelled as EPU articles.

3.3 EPU Indices based on SBERT Embeddings

The third approach applies Transformer based text embeddings (see Subsection 2.2) to identify articles that relate to EPU. To train SBERT articles' embeddings, Python's implementation of SBERT, namely *Sentence-Transformers*, and one of the pre-trained models, *distiluse-base-multilingual-cased-v2*, are used.[6] Due to computing capacity limits, the default setting of 128 tokens (word pieces) is adopted to extract the multilingual SBERT articles representations.

To construct the EPU time series, for each country, cosine similarities between all the articles' embeddings and the EPU embedding are calculated. The labeling process of the single articles is analogous to the previous approach: the articles whose embeddings are nearest to the EPU embedding as measured by the 90% percentile are labelled as EPU articles.

3.4 EPU Indices based on Multilingual Contextualized Topic Modeling

Finally, a novel language agnostic topic modeling technique introduced by Bianchi et al. (2021) called zero-shot cross-lingual topic modeling is applied (see Subsection 2.3). Thereby, a topic model is trained based on German SBERT articles embeddings and the topic distributions for Russian and Ukrainian articles are then inferred based on the corresponding articles embeddings. To identify EPU related topics, cosine similarity values are calculated between the EPU embedding and the topic embeddings.[7] In the current work, five topics the embeddings of which are nearest to the EPU embedding are labelled as EPU topics. The relative importance of each of these EPU related topics is calculated on a monthly basis. These relative importance time series for each of these topics over the considered time period can be directly used as an approximation for the EPU time series. Alternatively, one can use the average relative frequency of the the five identified EPU topics as an EPU time series. Overall, this approach reveals six EPU time series for each country.

[5] This threshold was determined in an iterative way. It can be also set to 95%, 99% etc.

[6] The complete documentation and examples are available on https://www.sbert.net/index.html.

[7] Topic embeddings represent the sum of the 100 word vectors corresponding to the most frequent topic words.

To distinguish between the proposed indices, these will be referred to as follows:

- *dic*1: a dictionary based approach that uses pre-trained word embeddings to identify three term sets related to the components of EPU concept.
- *dic*2: a dictionary based approach that uses pre-trained word embeddings to identify a compound EPU related term set.
- *art_emb*: an approach that constructs articles' embeddings based on the word embeddings corresponding to the articles' words.
- *art_sbert*: an approach that uses Transformer to learn articles' embeddings.
- $MCTM\_\{k\}\_Topic$: an approach that uses MCTM. k can stand for the topic number 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.

4 Evaluation

As the used text representation techniques are fully unsupervised, there are no standard way to evaluate the resulted text representations and, consequently, the resulted EPU indices. However, the constructed indices are believed to capture some relevant information about the general economic policy uncertainty. For this reason, I further analyse the constructed indices within VAR models to shed more light on the link between EPU and economic activity. In the analyses presented in the following, 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).

The data for the analyses come from the Federal Statistical Office of Germany[8] for Germany, from the Russian Federal State Statistic Service[9] for Russia, and from the State Statistics Service of Ukraine[10] for Ukraine. The data for Germany and Russia span the period from January 2000 to September 2020, for Ukraine - from January 2007 to September 2020.

In order to model the relationships between the EPU indices and economic activity, I estimate several two-dimensional VAR models for each country, each containing one of the constructed indices and the corresponding industrial production index. In general, VAR models consider all variables as endogenous and allow to account for possible interdependencies between them, which is especially important in the current case. First, one can argue that increased uncertainty level probably leads to a decrease in industrial production. Second, a drop in industrial production can cause an increase in economic related

[8]The data are available under https://www.destatis.de/DE/Service/OpenData/konjunkturindikatoren-csv.html (last accessed on 23.03.2021).

[9]The data are available under https://www.gks.ru/bgd/free/b00_25/IssWWW.exe/Stg/d000/i000850r.htm (last accessed on 23.03.2021). The values of the index for 2020 are available to the base year 2018. Therefore, the data for 2020 were recalculated to the base year 2010 as the rest of the data.

[10]The data are available under http://www.ukrstat.gov.ua/ (last accessed on 23.03.2021).

uncertainty. All models include seasonal dummies. The lag number is chosen according to the Akaike information criterion.

5 Results

In this section, first I present the resulted EPU indices. Then, the results of VAR analysis are presented.

5.1 EPU Indices

This subsection presents some of the resulted EPU time series and qualitatively analyses them.

Figure 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 figure of the EPU in Russia increases between 2004 and 2005 as well as in the period from 2014 and 2018. These peaks could be explained by the Orange Revolution in neighbouring 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.

Figure 3 shows the resulted indices based on the aggregated articles embeddings described in Subsection 3.2. One can observe substantial differences in the EPU level in the considered countries over the time. For example, while the EPU index increases in Germany in the period from the end of 2002 to the end of 2003, the figure of the EPU index in Russia drops in that period of time. It could be explained by the Iraq War and the reform course of the German government (Agenda 2010). At that time in Russia, the president Wladimir Putin declared the Second Chechen War to be over, which might have caused the drop in the uncertainty index in the mid-2003. However, the situation deteriorated again as the Chechen bombs explored in Moscow at the end of 2003. In following years, both countries witness a rise in the EPU index. This development could be associated with different domestic and foreign events like the change of the government in Germany, new developments in the Iraq War, terrorist attacks in London, natural disasters in Asia, USA, and Pakistan. In the period from the end of 2008 to the end end of 2009, all the countries experience an increase in EPU that is probably associated with the global financial crisis. Interestingly, as of mid-2018, the EPU time series for Germany and Russia move closely to each other and show an upward trend, while the figure of the EPU in Ukraine fluctuates around comparatively low values between 0.05 and 0.06.

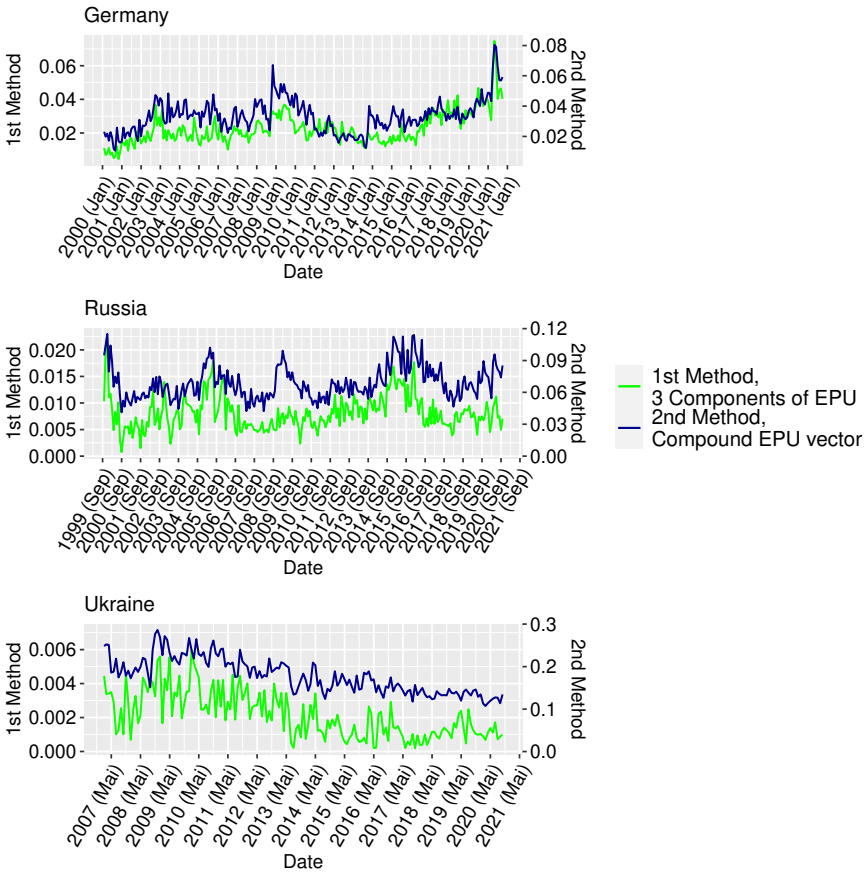

Figure 2: *dic*1 and *dic*2 EPU Indices

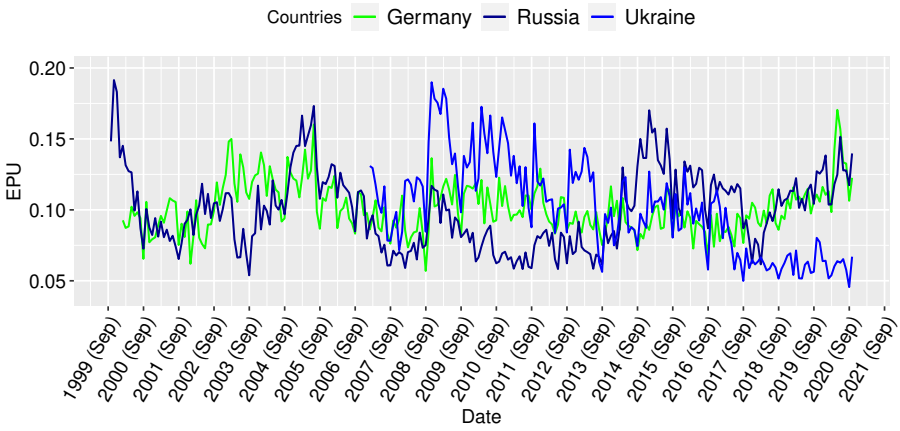

Figure 3: *art_emd* EPU Indices

Figure 4 shows the *art_sbert* indices for all three countries. In comparison to Figure 3, there are some noticeable differences in the time series' dynamics. For example, the EPU time series for Ukraine and Russia have more interdependencies what seems plausible due to the historical background, the geographical proximity, and the political entanglements of these countries.

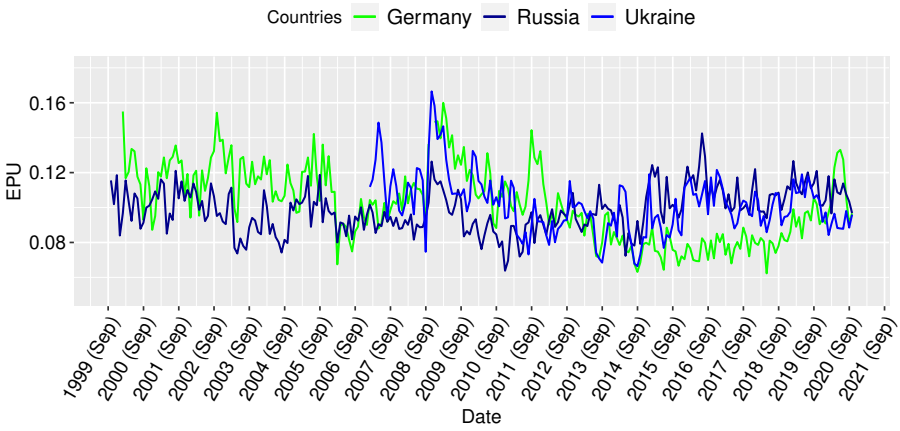

Figure 4: *art_sbert* EPU Indices

Finally, based on the results of the multilingual topic modeling described in Subsection 3.4 five EPU related topics are identified in the current application. These are presented in Figure 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 are presented in Appendix B.

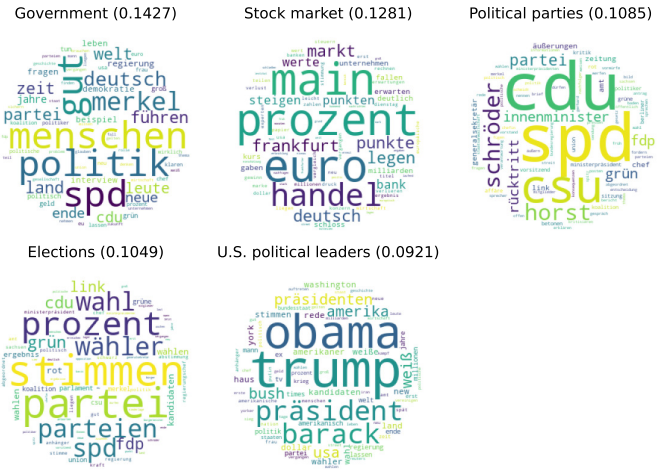

Figure 5: MCTM with 40 Topics: EPU Related Topics

5.2 VAR Models

For each country, 10 two-dimensional VAR models with seasonal dummies are estimated, each including one of the constructed EPU indices and the corresponding industrial production index.[11] The first set of analysis is dedicated to Granger causality tests. Granger causality tests allow to examine whether the lagged values of the variables in a VAR model actually affect each other. As the focus of the current work is on examining the impact of EPU on economic activity, especially the results of the null hypothesis "EPU does not Granger cause Industrial Production Index" are of interest. According to Granger causality tests, the following EPU indices led to significant results at least in one country: *dic*1 (Ukraine), *dic*2 (all countries), *art_emb* (Germany, Ukraine), *art_sbert* (Germany), $MCTM\_16\_Topic$ (all countries).[12] The results of Granger causality test are summarized in Appendix C.

To further investigate the link between EPU and economic activity, the reaction of a country's economic activity on shocks in the EPU was analysed

[11] 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 VAR models.

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

using Impulse Response Function (IRF). A close look is taken on the *dic*2 EPU index, as this index has proved to be Granger causal to economic activity in all the considered countries. Figure 6 illustrates the responses of the industrial production indices to an *dic*2 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 IRF in Ukraine is similar but not significant over the entire period.

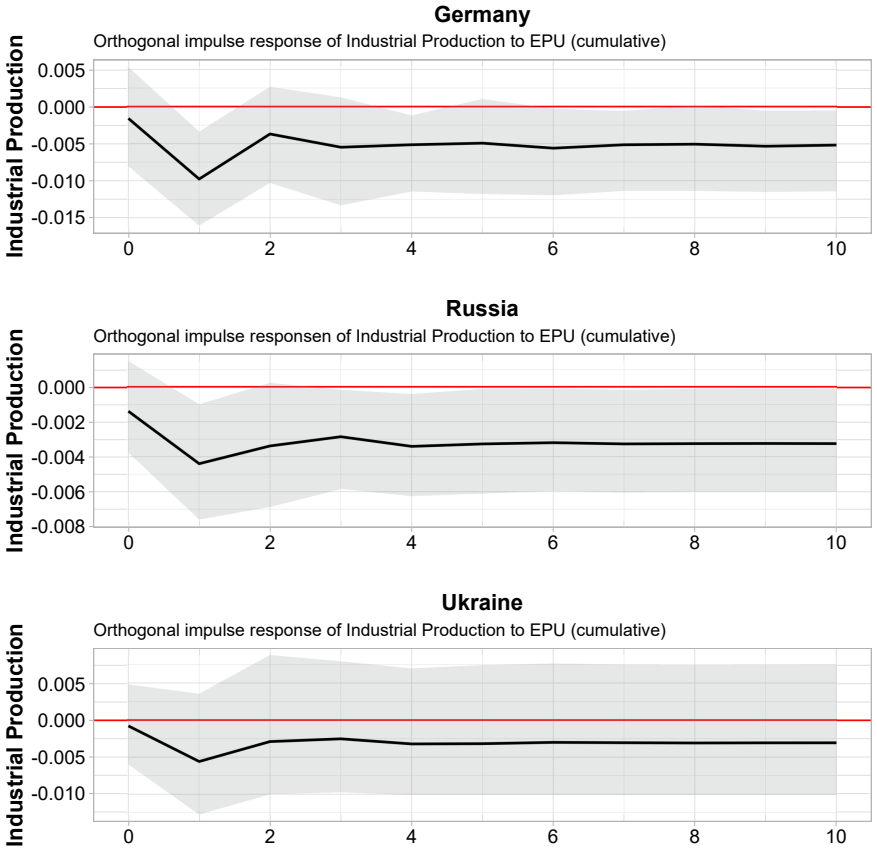

Figure 6: IRFs: *dic*2 EPU Index →Industrial Production Index

6 Conclusions

The current paper has focused on text-as-data applications in economics. The main goal of this paper was to construct uncertainty indices for different countries, namely Germany, Russia, and Ukraine, using multilingual text representation techniques.

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. This is an important insight for at least three reasons. First, it provides empirical evidence that even relatively simple methods could be successfully used to construct indices that can explain the economic activity of a country. Second, it speaks in favor of the usage of multilingual word embeddings in order to identify relevant EPU related terms in different languages. Third, while the process of creating relevant term sets has cost Baker et al. (2016) about two years and required experts with corresponding language knowledge and economic expertise to develop term sets for other countries, the proposed method here needs almost no supervision.

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.

The empirical findings of the current work indicate the usefulness of the proposed approaches to construction of EPU indices. Future research can be extended among the following lines. First, the proposed approaches can be improved by, for example, using more tokens in art_sbert approach and using some labelled EPU related articles as reference articles. Second, the presented approaches could be extended by other NLP techniques such as sentiment analysis. Third, the proposed approaches could be also used in other contexts, for example, by replacing the term "economic policy uncertainty" with "unemployment", "investment", "innovation" and obtaining specific news based indices describing the corresponding concepts. Finally, the combination with information sources like social media blogs, comments etc. might lead to more valid indices.

Appendix A EPU Terms

Economic Terms	"economic" or "economy"
Uncertainty Terms	"uncertainty" or "uncertain"
Policy Terms	"Congress", "deficit", "Federal Reserve", "legislation", "regulation" or "White House"

Table A1: EPU Terms by Baker et al. (2016)

Economic Terms	"wirtschaft" or "wirtschaftlich"
Uncertainty Terms	"unsicher" or "unsicherheit'
Policy Terms	"steuer" or "wirtschaftspolitik" or "regulierung" or "regulierungs" or "ausgaben" or "bundesband" or "ezb" or "zentralbank" or "haushat" or "defizit" or "haushaltsdefizit"

Table A2: EPU Terms for Germany by Baker et al. (2016)

Economic Terms	"экономика" (economy)
Uncertainty Terms	"неопределённый" (uncertain) or "неопределённость" (uncertainty)
Policy Terms	"политика" (policy), "налог" (tax)

Table A3: EPU Terms for Russia by Baker et al. (2016)

Economic Terms	"wirtschaftliche" (0.7934), "ökonomische" (0.7865), "wirtschaftspolitische" (0.7746), "wirtschaftsentwicklung" (0.7525), "volkswirtschaftliche" (0.7461), "gesamtwirtschaftliche" (0.7239), "wirtschaftswachstums" (0.7187), "marktwirtschaftliche" (0.717), "weltwirtschaft" (0.7098)
Policy Terms	"policy" (0.668), "informationspolitik" (0.6379), "richtlinienkompetenz" (0.6344), "ausländerpolitik" (0.6234), "grundsatzentscheidungen" (0.6161), "richtlinien" (0.6145), "neutralitätspolitik" (0.612), "ordnungspolitik" (0.6107), "wirtschaftspolitik" (0.6106), "rechtspolitik" (0.603), "beschäftigungspolitik" (0.5976), "industriepolitik" (0.5962), "migrationspolitik" (0.5939), "währungspolitik" (0.5934), "deutschlandpolitik" (0.5922), "regierungspolitik" (0.5921), "gesellschaftspolitik" (0.5914), "verteidigungspolitik" (0.5913), "politik" (0.5907)
Uncertainty Terms	"unsicherheit" (0.7323), "ungewissheit" (0.6751), "messunsicherheit" (0.6411), "eintrittswahrscheinlichkeit" (0.6298), "zufälligkeit" (0.6138), "wahrscheinlichkeiten" (0.6022), "zeitlichkeit" (0.5974), "messbarkeit" (0.5973), "voraussagen" (0.5956), "unbestimmtheit" (0.5943), "wahrscheinlichkeitsverteilung" (0.5877), "ausfallwahrscheinlichkeit" (0.5875), "erwartungswerte" (0.5846), "relativität" (0.5846), "berechenbarkeit" (0.5792), "bestimmtheit" (0.5787), "allgemeingültigkeit" (0.578)

Table A4: EPU Terms and their Cosine Similarity Values for Germany based on *dic*1 approach

Economic Terms	"экономического" (economic, 0.7818), "экономик" (economies, 0.7466), "макроэкономических" (macroeconomics, 0.7326), "социально" (socially, 0.6664), "рыночных" (market, 0.6537)
Policy Terms	"небюрократия" (non-bureaucracy, 0.5588), "правила" (rules, 0.5532), "законодательства" (legislation, 0.547), "политике" (policy, 0.5465), "рекомендаций" (recommendations, 0.532), "ужесточении" (tightening, 0.5308), "неприемлемы" (unacceptable, 0.5215), "лоббизм" (lobbying, 0.5181), "ивп" (iwp, 0.5164), "недопустимости" (inadmissibility, 0.5147), "правилах" (rules, 0.5134)
Uncertainty Terms	"неопределённость" (uncertainty, 0.6886), "относительность" (relativity, 0.6237), "определённость" (certainty, 0.618), "противоречивость" (inconsistency, 0.5964), "неясность" (ambiguity, 0.5875), "неуверенность" (uncertainty, 0.5655), "неизбежность" (inevitability, 0.5632), "адекватность" (adequacy, 0.5543), "закономерность" (regularity, 0.5535), "субъективность" (subjectivity, 0.5494), "обоснованность" (validity, 0.5483), "согласованность" (consistency, 0.5482)

Table A5: EPU Terms and their Cosine Similarity Values for Russia based *dic*2 approach

Economic Terms	"економіч" (economic, 0.7607), "економічна" (economic, 0.7348), "економік" (economies, 0.7208), "макроекономічної" (macroeconomics, 0.6612), "економіці" (economics, 0.645), "підприємництва" (entrepreneurship, 0.6332), "зовнішньополітична" (foreign policy, 0.6292), "зовнішньоекономічна" (foreign economic, 0.6281)
Policy Terms	"політики" (politics, 0.6205), "політикуму" (politicum, 0.5964), "зовнішньополітичний" (foreign policy, 0.5707), "законодавства" (legislation, 0.5596), "етнополітики" (ethnopolitics, 0.5475), "геополітики" (geopolitics, 0.5341), "політизації" (politicisation, 0.5284), "політиці" (politics, 0.5268), "мінагрополітики" (agricultural politics, 0.5228)
Uncertainty Terms	"невизначеність" (uncertainty, 0.6846), "невизначеності" (uncertainty, 0.6778), "визначеність" (certainty, 0.6074), "ймовірності" (probability, 0.5939), "імовірності" (probability, 0.5836), "визначеності" (certainty, 0.5833), "непередбачуваність" (unpredictability, 0.5774), "спостережуваного" (observable, 0.5761), "невизначеного" (uncertain, 0.574), "вірогідності" (probability, 0.5724), "неясність" (ambiguity, 0.5677), "взаємозалежність" (interdependence, 0.5647), "передбачуваного" (predictable, 0.5603), "нестабільність" (instability, 0.5573), "ймовірнісні" (probabilistic, 0.5545), "відносність" (relativity, 0.5545), "суперечливість" (inconsistency, 0.5542), "імовірність" (probability, 0.5523)

Table A6: EPU Terms and their Cosine Similarity Values for Ukraine based on *dic*1 approach

Germany	"wirtschaftspolitische" (0.7891), "marktwirtschaftliche" (0.748), "industriepolitik" (0.7362), "beschäftigungspolitik" (0.7326), "währungspolitik" (0.7243), "konjunkturpolitik" (0.724), "fiskalpolitik" (0.7188), "gesellschaftspolitik" (0.7186), "entwicklungspolitik" (0.7123), "steuerpolitik" (0.7107), "ökonomische" (0.7105), "volkswirtschaftliche" (0.7105), "finanzpolitik" (0.7093), "regierungspolitik" (0.7054)
Russia	"экономического" (economic, 0.7026), "макроэкономических" (macroecnomic, 0.6971), "экономик" (economics, 0.6913), "либерализации" (liberalisation, 0.6514), "внешнеполитическая" (foreign policy, 0.6466), "внешнеэкономическая" (foreign economic, 0.6464), "предпринимательства" (entrepreneurship, 0.6431)
Ukraine	"зовнішньополітична" (foreign policy, 0.689), "економіч" (economic, 0.6869), "економічна" (economic, 0.6793), "економік" (economics, 0.6734), "макроекономічної" (macroeconomics, 0.6606), "політика" (policy, 0.6547), "зовнішньоекономічна" (foreign economic, 0.6389)

Table A7: EPU Terms and their Cosine Similarity Values for Germany, Russia, and Ukraine based on *dic*2 approach

Appendix B MCTM

In Appendices B.1, B.2, and B.3, the corresponding topic time series are shown. The horizontal red dashed lines represent the average topic probability (2.5%) if all of the 40 topics were equally distributed. In Germany, for example, the time series of stock market topic experience a considerable increase up to 9% in the period from the end of 2000 to 2002 that could be explained by the dot-com bubble. Time series of `elections` topic constantly fluctuates throughout the considered time period with some considerable spikes around major political events like parliamentary elections (2002, 2005, 2009. 2013). Finally, major spikes of the figure for the relative importance of `U.S. political leaders` are probably associated with the presidential elections in the USA.

At this point, it should be highlighted that BoW representations of the German texts are only used to visualize topics that were trained based only on SBERT embeddings. It means that topics were trained based on semantic similarity of the underlying texts. While these topic representations can be meaningfully interpreted for the German dataset, as it was used for training, one should be careful when describing inferred topic distributions for Russia and Ukraine. For example, Russian or Ukrainian articles assigned to T`political parties` topic most probably do not only report on German political parties but also on domestic key political players. The analysis of the Ukrainian articles assigned to this topic revealed that they mainly report on the key events in Verkhovna Rada (unicameral parliament of Ukraine) and major political Ukrainian parties and coalitions. With regard to other EPU related topics, it has been observed that articles with the highest probability for `government` and `elections` topics discuss political themes in general, for `stock market` topic - movements of stock markets at home and worldwide, for `U.S. political leaders` topic - political leaders of the USA.

B.1 Germany

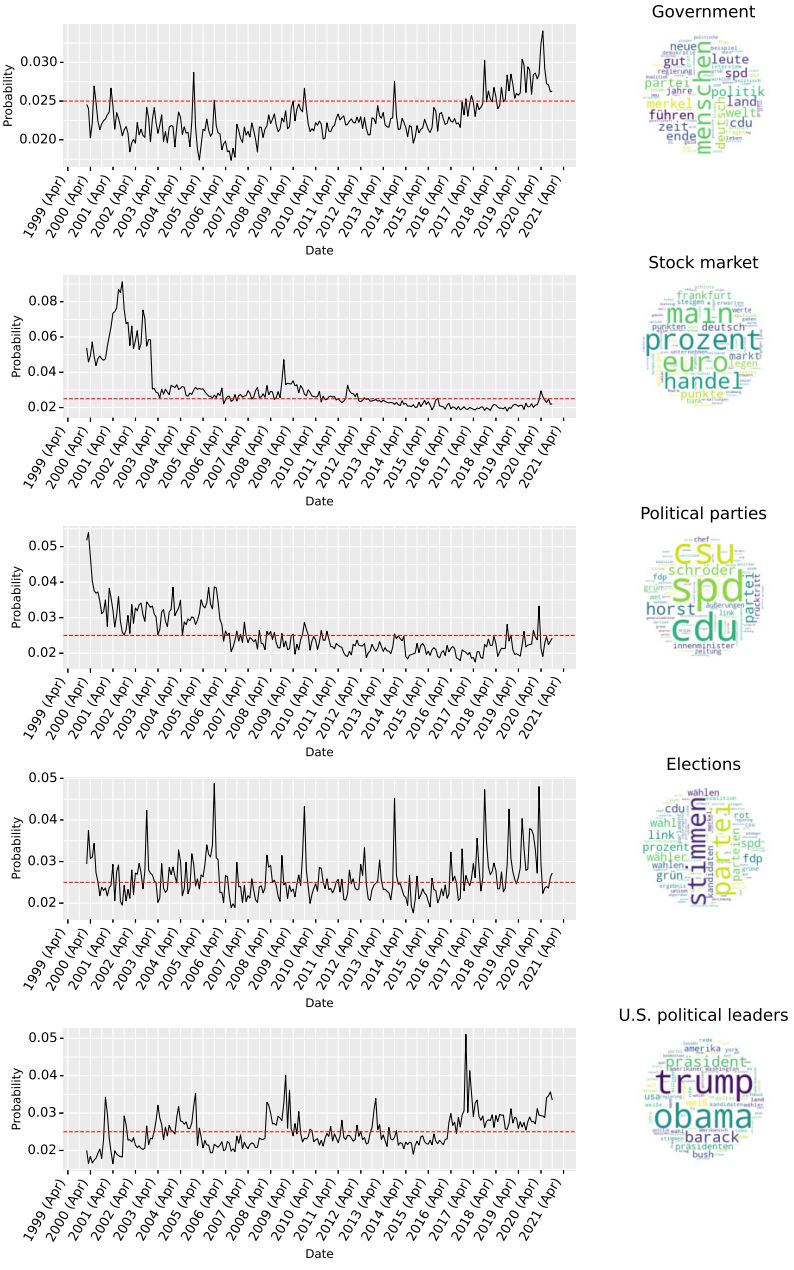

Figure B1: Time Series of Selected EPU Topics in Germany

B.2 Russia

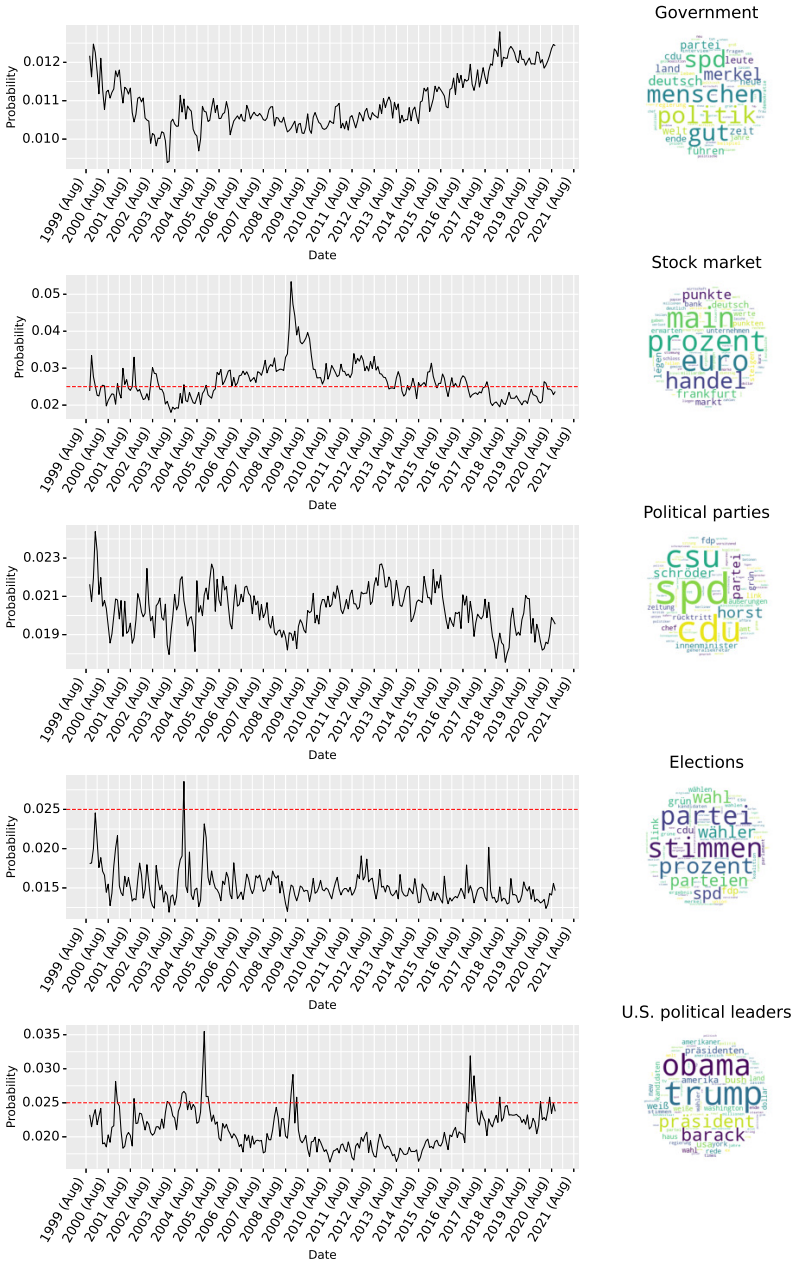

Figure B2: Time Series of Selected EPU Topics in Russia

B.3 Ukraine

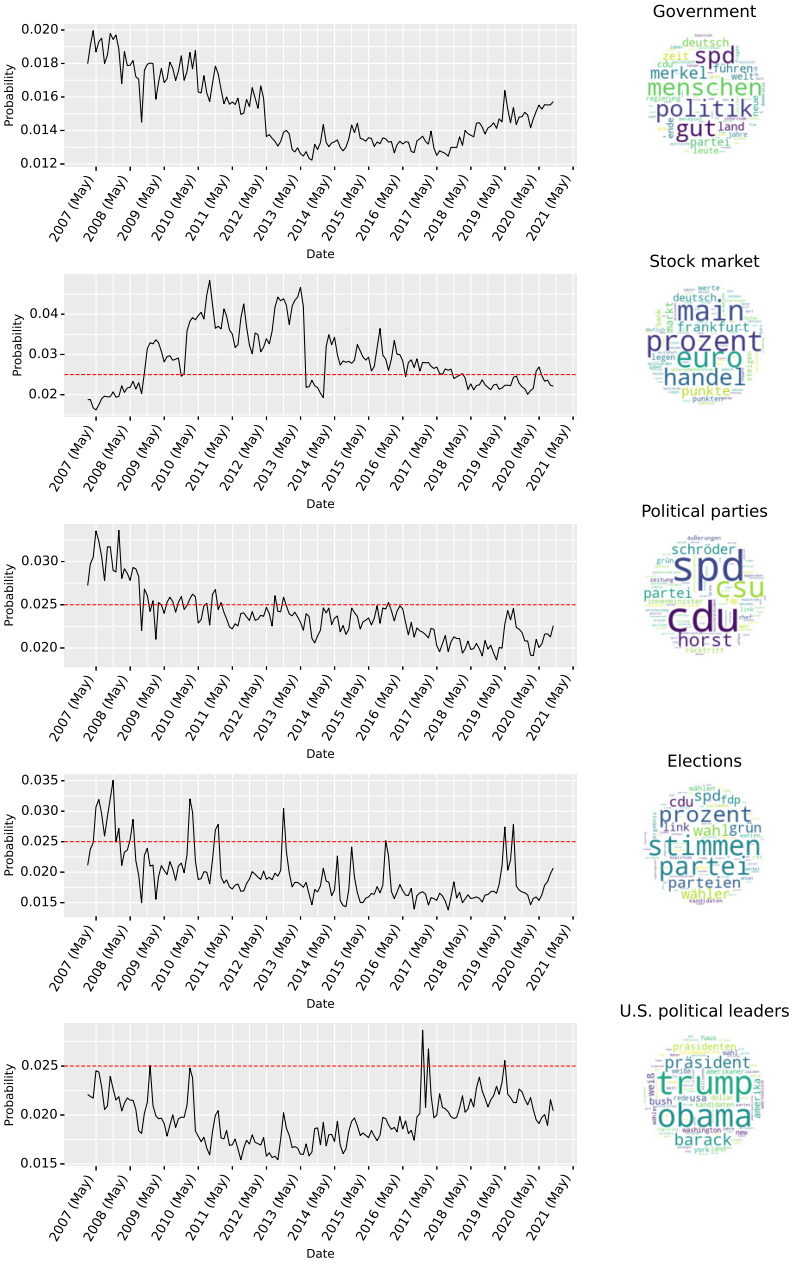

Figure B3: Time Series of Selected EPU Topics in Ukraine

Appendix C Granger Causality Tests

Country		Dictionary Based		Aggregated Document Embeddings	SBERT	MCTM					
		separated terms	combined terms			Topic 10	Topic 36	Topic 4	Topic 0	Topic 12	Combined
Germany	EPU →IND	0.1153*	**0.02374***	**0.07272***	**0.009405**	0.196*	**0.05637***	**0.08755'**	0.7107'	0.2807	0.1853*
	IND →EPU	0.9654*	**0.04103***	0.2034*	**0.01072**	0.583*	0.5979*	0.3958'	0.5694'	0.4474	**0.03061***
Russia	EPU →IND	0.2496"	**0.06071"**	0.254'	0.2504'	**0.01369***	**0.008587'**	0.164"	**0.003431'**	0.7935'	0.1468'
	IND →EPU	0.7121"	0.611"	0.6064'	0.2818'	0.5302*	0.7798'	0.3228"	0.2363'	0.6583'	0.4365'
Ukraine	EPU →IND	**0.08736***	**0.08407***	**0.00461***	0.1311*	0.2746*	**0.06555***	**0.02054***	0.5193*	0.1874	0.2617*
	IND →EPU	0.75381*	0.2311*	0.1903*	0.9265*	0.809*4	0.9312*	0.07189*	0.443*	0.3294	0.4316*

*: The residuals of the model are heteroscedastic.
": The residuals of the model are autocorrelated.
': The residuals of the model are heteroscedastic and autocorrelated.
The values in bold are significant at 10% significance level.
Dictionary based: approach presented in Subsection 3.1. *Separated terms* refers to the first method, where three terms sets are defined
that describe three components of the EPU concept. *Combined terms* refers to the second method, where a single term set is defined to describe EPU.
Aggregated Document Embeddings: approach presented in Subsection 3.2. Articles are represented as the sum of constituent words (fastText embeddings).
SBERT: approach presented in Subsection 3.3. SBERT embeddings are used to represent articles' texts.
MCTM: approach presented in Subsection 3.4. The uncovered Topics 10, 36, 4, 0, and 12 have been identified as EPU topics. Additionally, a combined topic has been constructed.
IND: the growth rates of the industrial production indices as they were used in VAR models.
EPU: the growth rates of the corresponding constructed EPU indices as they were used in VAR models.

Table C8: Granger Causality Test Results (p-values)

References

Algaba, A., S. Borms, K. Boudt, and J. van Pelt. 2020. The economic policy uncertainty index for flanders, wallonia and belgium. *SSRN Electronic Journal*: BFW digitaal/RBF numérique 2020/6 .

Azqueta-Gavaldón, A. 2017. Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters* 158: 47–50 .

Baker, S.R., N. Bloom, and S.J. Davis. 2016. Measuring economic policy uncertainty*. *The Quarterly Journal of Economics 131*(4): 1593–1636 .

Bianchi, F., S. Terragni, D. Hovy, D. Nozza, and E. Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. *CoRR* abs/2004.07737. https://arxiv.org/abs/2004.07737 [cs.CL].

Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM 55*(4): 77–84. https://doi.org/10.1145/2133806.2133826 .

Čižmešija, M., I. Lolić, and P. Sorić. 2017. Economic policy uncertainty index and economic activity: what causes what? *Croatian Operational Research Review 8*(2): 563–575 .

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. 2017. Word translation without parallel data. *CoRR* abs/1710.04087. https://arxiv.org/abs/1710.04087 .

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova 2019, June. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Ghirelli, C., J.J. Pérez, and A. Urtasun. 2019. A new economic policy uncertainty index for spain. *Economics Letters* 182: 64–67 .

Kapfhammer, F., V.H. Larsen, and L.A. Thorsrud 2020, December. Climate Risk and Commodity Currencies. Working Papers No 10/2020, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Kaveh-Yazdy, F. and S. Zarifzadeh. 2021. Measuring economic policy uncertainty using an unsupervised word embedding-based method.

Lenz, D. and P. Winker. 2020, 01. Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE 15*(1): 1–18. https://doi.org/10.1371/journal.pone.0226685 .

Lolić, I., P. Sorić, and M. Logarušić. 2022, aug. Economic policy uncertainty index meets ensemble learning. *Comput. Econ. 60*(2): 401–437. https://doi.org/10.1007/s10614-021-10153-2 .

Manela, A. and A. Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics 123*(1): 137–162 .

Miranda-Belmonte, H.U., V. Muñiz-Sánchez, and F. Corona. 2023. Word embeddings for topic modeling: An application to the estimation of the economic policy uncertainty index. *Expert Systems with Applications* 211: 118499. https://doi.org/https://doi.org/10.1016/j.eswa.2022.118499 .

Perić, B.Š. and P. Sorić. 2018. A note on the "economic policy uncertainty index". *Social Indicators Research 137*(2): 505–526 .

Reimers, N. and I. Gurevych 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)*, Hong Kong, China, pp. 3982–3992. Association for Computational Linguistics.

Reimers, N. and I. Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR* abs/2004.09813. https://arxiv.org/abs/2004.09813 [cs.CL].

Xie, F. 2020. Wasserstein index generation model: Automatic generation of time-series index with application to economic policy uncertainty. *Economics Letters* 186: 108874 .