Mendel University in Brno

MENDELU Working Papers
in Business and Economics
89/2023

A Robustness Analysis of Newspaper-based Indices

Roman Valovič, Daniel Pastorek

MENDELU Working Papers in Business and Economics
Research Centre
Faculty of Business and Economics
Mendel University in Brno
Zemedelska 1, 613 00 Brno
Czech Republic
http://vyzc.pef.mendelu.cz/en
+420 545 132 605

Citation

Valovič, R. and Pastorek, D. (2023). A Robustness Analysis of Newspaper-based Indices. *MENDELU Working Papers in Business and Economics* 89/2023. Mendel University in Brno. Cited from: http://ideas.repec.org/s/men/wpaper.html

Abstract

Roman Valovič, Daniel Pastorek: **A Robustness Analysis of Newspaper-based Indices**

In this paper, we subject the methodology for newspaper-based indices to several tests of robustness, to address the potential problems of this proposed approach. Firstly, we examine the strong dependency between the selected keywords and the entered query. We do this using state-of-the-art language models, such as BERT, to automatically select relevant articles to build the index. Secondly, we propose that the weighting of articles partly allows for the control of the context of the articles and potential errors in the incorrect identification of articles, which leads to more stable index results. Finally, we track composition changes in newspaper articles, which have been evolving over time. The implications of these tests may be of interest to the users of these indices as well as suggesting a future direction for this approach.

Key words

newspapers, economic-policy uncertainty, EPU index, NLP, text-mining, similarity search

JEL: C80, D80, E80, E66

Contacts

Roman Valovič, Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemedelska 1, 613 00 Brno, Czech Republic, e-mail: roman.valovic@mendelu.cz.

Daniel Pastorek, Department of Finance, Faculty of Business and Economics, Mendel University in Brno, Zemedelska 1, 613 00 Brno, Czech Republic, e-mail: daniel.pastorek@mendelu.cz

Acknowledgements

This outcome was supported by the Internal Grant Schemes of Mendel University in Brno, registration no.: CZ.02.2.69/0.0/0.0/19_073/0016670, funded by the ESF.

1 Introduction

Estimating the impact of uncertainty, as one of the most highly debated macroeconomic factors that leads to a significant decline in economic activity, has been the subject of fertile research over recent years. Identifying uncertainty, that is not directly measurable, poses a significant challenge and several proxies for uncertainty have been created. While some approaches use a data-rich model of the volatility of the forecast errors of a large number of economic indicators (e.g., Jurado et al. (2015)), others use more traditional proxies as surveys of forecast errors, or use stock market volatility indices such as VIX, the greatest degree of attention in econometric studies has been paid to the relatively simple (albeit inventive) newspaper-based approach devised by Baker et al. (2016) [1]. It is the simplicity of this approach, which tracks the frequency of newspaper articles which contain a combination of selected keywords, that has made this variable easy to use in empirical studies, but it has also raised several relevant questions about the robustness of these results, which have not yet been answered.

In this paper, we seek to test the robustness of the current newspaper-based methodology with respect to a number of potential problems we are aware of that come from this methodological approach. Firstly, the articles are usually obtained from newspaper databases where the resultant set of articles is strongly dependent on the selected keywords and the query used (usually a logical expression that the article must meet). However, such a search does not take into account polymorphism or possible synonyms of the keywords used in the query. Also, the keyword selection is highly subjective. Secondly, the methodological approach proposed by Baker et al. (2016) does not allow us to distinguish between articles that discuss a true in-depth uncertainty in economy and articles that only brush over the surface of the issue[2]. Thus, the question we seek to answer is whether all uncertainty peaks are caused by an increase in uncertainty. Thirdly, the news-based indices show notable trends that are not as clear in other proxies for uncertainty. Therefore, we investigate to what extent these trends are driven by uncertainty itself or are due to changes in the composition of journals or the individual preferences of editors over time.

These exercises are conducted to examine how robust the results remain after we control for these potential problems. Additionally, as we subject our data to detailed analyses, it allows us to present the distribution of keywords in the articles in greater detail. To compare the results of the original approach and our tests, we use the most common economic-policy uncertainty index from Baker et al. (2016) as a benchmark. Our results highlight how sensitive the current methodology is to these proposed issues. Furthermore, we also provide a detailed analysis of the robustness of this approach to users of such indices, suggesting potential approaches to optimize the methodology.

Results in our paper show that after a large and gradual expansion of the number of keywords (including polymorphisms and possible synonyms), the correlation between such an index and the original approach starts to decrease slightly. Simultaneously, the gradual addition of words causes the smoothing of the index. We find that by using individually weighted articles, we can obtain more stable index parameters and higher correlations. In addition, by weighting articles, we reduce the error rate of the methodology and partially control for the context of the articles. We also suggest that through a focus on articles that are categorized according to their content, changes in trends might be better distinguished from spurious changes caused by po-

[1] We describe this methodology in more detail in chapter 2. This methodology was proposed by Baker et al. (2016), but many other indices using a similar approach are also available. For a partial overview see `https://policyuncertainty.com`

[2] We assume that an article that is entirely devoted to an economic issue should receive more weight than an article that addresses it in a marginal context. Also, the random distribution of words in a text can result in the fulfilment of condition and will carry the same weight as an economically focused article.

tential changes in the composition of journals or changes in the individual preferences of editors over time. The results of these tests can be generalized for the various newspaper-based indices under consideration. Furthermore, these results can serve as an indication of the direction of future developments in this methodological approach.

The paper is structured as follows. Section 2: the newspaper-based methodology and currently proposed approaches. Section 3: an overview of the data and methods. Section 4: the results of the analyses. Section 5: conclusions. The paper ends with an appendix that contains additional information about the data used.

2 The current newspaper-based approaches

Printed (digitized) or electronic daily newspapers serve as the primary data source from which the authors selected articles that met their criteria, formulated as follows. An article is considered relevant if it contains at least one word from each of the 3 defined groups of keywords:

- Economy (E): "economy", "economic",
- Uncertainty (U): "uncertainty", "uncertain",
- Policy (P): "congress", "deficit", "Federal Reserve", "legislation", "regulation", "White House".

This can be formally described as below:

$$e = \begin{cases} 1 & \text{if } w_{iE} \in A_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$u = \begin{cases} 1 & \text{if } w_{iU} \in A_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$p = \begin{cases} 1 & \text{if } w_{iP} \in A_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where w_i is a search term for group E, U and P respectively and A_j is j-th article of the corpus. Article A_j s relevant if the result of C is true.

$$C : e > 0 \wedge u > 0 \wedge p > 0 \tag{4}$$

These search terms can only be used for English-language newspapers and for a specific country or region. As an example, for German daily newspapers, alternative sets with different terms ("wirtschaft", "wirtschaftlich", "wirtschaftspolitik" etc.) have to be used. However, the original approach does not take into account polymorphism nor possible synonyms of the keywords used in the search query. That is, the only articles retrieved are those with an absolute match between the search term and the keyword in the article. This can lead to the non-retrieval of relevant articles, even though they contain semantically similar words, thus the method proposed by Baker et al. (2016) can produce a number of false negatives. Several alternative approaches have been proposed in recent years to improve this widespread and accepted methodology. Miranda-Belmonte et al. (2023) represented text using GloVe embeddings, to cluster articles

according to their topic. The EPU index was generated by counting the number of articles assigned to each topic. However, the number of clusters (i.e., topics) must be defined manually, which is difficult when the domain is unknown. Vargas-Calderón and Camargo (2019), noted, that a too low number of clusters causes internal heterogeneity (documents that contain different topics are placed in a common cluster), while a high value causes external homogeneity (documents with the same or similar topics are placed in different clusters). Azqueta-Gavaldon et al. (2020) used LDA allocation to represent each article as a set of words, which represent a specific topic. The resulting set of topics are then aligned with the original Baker categories. However, this process must be done manually, requiring human supervision. Tobback et al. (2018) used SVM classifiers to predict whether articles were EPU-relevant or non-EPU-relevant in order to build the Belgian EPU index. Since they used tf-idf (Weiss, 2015) for text representation to feed the SVM classifier, they were unable to identify the semantic relationship between words. Moreover, manual labelling of training data is necessary.

3 Data and Methods

In this section, we will present our data and the approaches used to test the proposed potential problems we are aware of within this methodology in more detail. Firstly, to keep our results as consistent as possible, we follow the Baker et al. (2016) approach, and collect electronic articles from the same data sources. The authors have indices available for various countries, but we will focus on the data for Germany using its own newspaper archives, as per the EPU index. We do so because for most of the other indices authors using intermediary databases, may use data that might not be fully consistent with the publishers' data. Either as a consequence of the method used for data retrieval or due to issues with individual filters used in the selection of articles in these databases, or other issues such as multiplicity etc. More specifically, we use the German newspapers, Handelsblatt (containing 346,833 articles), and Frankfurter Allgemeine Zeitung (683,622 articles) from 2012 to mid-2022. Since we focus on newspapers published in electronic format, we had to limit the time scale to 2012 onwards due to gaps in the data and the uneven distribution of the number of articles in individual months before 2012.

3.1 Data preprocessing

Text data extracted from documents is unstructured, therefore it must be transformed into a structured form that allows machine processing. Preprocessing usually includes steps such as the removal of stop-words (words with no semantic meaning, prepositions, conjunctions, etc.), conversion to lowercase, and the removal of digits and special characters. Given that the nature of the task we are dealing with is word-level parsing, we can afford a more stringent level of preprocessing, since we do not need to preserve the contextual relations between words. In order to speed up the analysis of a large number of articles, our goal is to reduce the text, to the greatest extent possible, while preserving the relevant words. Another well-known preprocessing technique is stemming and lemmatization. Stemming is the process of truncating the suffixes or prefixes of words to a common base. For example, the words "economy" and "economic" can be stemmed to "econom" or "econ". The resulting words do not necessarily have to make sense, the goal is to reduce the corpus vocabulary as much as possible. Lemmatization converts words to their dictionary form, e.g., "economies" to "economy" (Weiss et al., 2015). Both approaches may partially solve the problem of the occurrence of synonyms ("economy" and "economic"), but it certainly cannot be considered to be a comprehensive solution. In our work we lemmatized the text.

3.2 Similarity search with BERT

To solve the problem of the identification of synonyms, we represent a text by so-called word vectors (or embeddings). This principle is based on the fact that text (words, sentences and possibly paragraphs) may be represented by a numerical vector, usually in a high-dimensional space (one hundred or more), such that semantically similar words are represented by a similar vector. Mathematically speaking, the lower the distance between vectors is, the more similar the words are. Cosine similarity was used as the distance metric. The concept of vector representation was introduced by Mikolov et al. (2016) with his word2vec algorithm and later improved by Bojanowski et al. (2017). The current state-of-the-art in the field of text representation are language models, the so-called Transformers, based on the BERT neural network (Vaswani et al., 2017).

In addition to the very precise embeddings produced by these models, many of them are multilingual. That is, one model can be used to encode text in different languages. Since training a custom model is computationally expensive, we used the readily available *bert base multilingual uncased* model (Reimers and Gurevych, 2019), which allows the encoding of text in 104 languages, making it usable for a wide range of newspapers and other sources. In other words, we can use a single model to analyse German, English, or Italian articles. This would not be possible using word2vec or Fasttext, as these legacy techniques are mono-lingual, meaning that each language requires its own specific model. Additionally, using BERT makes it easier to reproduce our approach as the model is freely available[3] and deterministic.

3.3 Definitions

In the following section, we describe the meaning of some of the expressions and variables that have been used in this work.

Search embeddings for each search term, formally V_{Ei}, V_{Ui}, V_{Pi} is the vector of the i-th search term of the group E, P and U respectively[4]. We used BERT to get the vector of each word in corpus.

Corpus dictionary, it holds all the unique words that occur in the text. In this paper, we denote the dictionary as D. The total number of unique words (dictionary size) is denoted by J.

Cosine similarity is a metric used to measure how similar vectors are to each other. As mentioned above, semantically close words also have close vectors. The cosine similarity calculation, measured by the angle of vectors is defined as follows:

$$Sim(V_y, V_y) = \frac{\sum_{i=1}^{n} V_{xi} V_{yi}}{\sqrt{\sum_{i=1}^{n} V_{xi}^2} \sqrt{\sum_{i=1}^{n} V_{yi}^2}} \tag{5}$$

Corpus vectors V_j represents the vector of the j-th word in the corpus dictionary D.

Using the definitions above, we can formulate a condition for semantically similar words, which states that words D_i and D_j are similar to each other if cosine similarity of theirs vectors (denoted as V) satisfies the threshold ST_{min}:

$$Sim(V_x, V_j) \geq ST_{min} \tag{6}$$

[3] See `https://huggingface.co/bert-base-multilingual-uncased`

[4] The groups E, P and U represent the categories Economy, Policy, and Uncertainty, as proposed by Baker et al. (2016) presented in section 2.

Thus, in our approach, we do not require a strict match between a particular search term and a word in the article. Instead, we evaluate whether the condition is satisfied based on mutual similarity. We can modify the condition defined in section 2 as follows. The analogy also applies to groups U and P. Condition C remains unchanged.

$$e = \begin{cases} 1 & \text{if } Sim(w_{iE}, w_{jA}) \geq ST_{min} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

It follows that the condition can even be met by articles that do not contain words that are identical to the search term. It is necessary to pay attention to the value of the parameter ST_{min}.

Similarity threshold ST_{min} is a quantitative expression of the semantic similarity of two words on a scale from 0 to 1. If it is set strictly (close to 1), no synonyms will be accepted and there is the risk that fewer relevant articles will be retrieved. On the other hand, too low a value results in semantically distant words being considered to be synonyms. Although the number of articles retrieved will be higher, a higher number of non-relevant articles will also be counted resulting in a higher level of bias. We provide examples of extended search word sets in Appendix B. Constraint C thus considers both user-specified words, but also derived words whose similarity is not less than the specified threshold ST_{min}.

3.4 Vector database

However, from an implementation point of view, this means that the similarity between each search term and each word in dictionary D needs to be calculated, which is computationally expansive and thus time-consuming. For this reason, we used the Milvus vector database. Milvus functions as a relational database but allows fast searches in the vector space, as each component of the vector is indexed in the structure. Thus, the database stores a dictionary D, and a corresponding precomputed embedding for each word. This can be used to obtain a set of similar words. In our approach, we used Milvus to automatically expand the search terms of each group - for each search term, the database returns all words whose cosine similarity is less than a given threshold ST_{min}. We then query the presence of the search term w_{iE}, w_{iU} and w_{iP} in article A_j by a simple set operation. This technique significantly shortens the search and evaluation of the corpus of articles.

3.5 Algorithm

Our proposed approach to retrieve the relevant articles to compile the EPU index can be summarized in the following steps:

1. Get embeddings of all the words in the corpus dictionary D and input search terms.
2. Calculate the cosine similarity between each input term for the Economy (E), Policy (P), and Uncertainty (U) groups and words from the dictionary D. Words whose similarity is higher than the ST_{min} threshold will extend the corresponding group (E, P, or U).
3. Iterate through all the articles and count the number of occurrences of the search terms for groups E, P and U in each article and evaluate condition C.
4. Generate the EPU index according to Baker (2016):
    (a) For each newspaper p: Compute relative number of articles of p that meet the condition C for each period t. Denote as X_{pt}.

(b) Divide X_{pt} by standard deviation of X_{pt}. Denote as STD_p

(c) Compute mean of STD_p of all newspapers in period t. Denote as M_t

(d) Compute mean M_t in all periods. Denote as M.

(e) Divide STD_p by STD_t and multiply by 100. Denote as **EPU**.

5. normalize the values using min-max normalization in order to rescale the values of EPU into the interval [0, 100] using equation 8.

$$EPU_{norm} = \frac{EPU - min(EPU)}{max(EPU) - min(EPU)} \cdot 100 \tag{8}$$

3.6 Article categories

In the online resources, the individual articles are categorized according to content. The categories vary from newspaper to newspaper, but in general we can find an overlap. Handelsblatt, for example, categorizes articles into Politik, Unternehmen, Technologie, Finanzen, Mobilität, Karriere, Arts & Style, Meinung and Video[5]. In the original index composition, the number of articles that meet the specific conditions for keywords are divided by the total number of all articles published in any specific month. However, this does not control for the composition change in newspaper articles over time, which naturally evolve. As an example, the transition from a high quantity of articles to longer articles that consider the topic in a greater depth or changes in the overall number of articles published in a specific category (e.g., a shift from technology-focused articles to more politically oriented news). These transitions affect the denominator of indices. However, in this case, the ratio change over time may rather reflect changes in the journal or the preferences of individual editors than long-term trends in the specific indices. Using data regarding economic-political uncertainty, the question arises, whether the levels of uncertainty over time are due to an increase in uncertainty or these transitions.

To answer this question, we first need to unify the article categories across all sources. We will use the Handelsblatt categories as a reference list, so we need to classify the Frankfurter Allgemeine Zeitung articles accordingly. This can be done manually, but it is laborious and in some cases the category of the article is not given. The researcher would then have to read the article and decide where to put it, which is unrealistic given the number of articles.

Therefore, we use the Handelsblatt dataset of articles, where the category of each article is also given. We use the data to train a model that is able to automatically assign the category to article based on its content. We used this model to categorize the Frankfurter Allgemeine Zeitung articles, so we obtained a uniform set of categories for both newspapers. This approach is applicable to all newspaper-based indices that use different newspapers as sources of data. For this purpose, we again used the BERT neural network with a fine-tuned classification layer; the results for several network configurations are shown in Table 1. We were able to score 79% in the F1-metric, with a dataset split of 80% for training, and 20% for testing. The best result was achieved using 5 epochs with the learning rate set to 0.0001. The learned model is capable of categorizing new articles coming from various sources. As can be seen, the selected multilingual language model can ensure high quality text embedding in German. We also tried to predict the category according to the article title alone, as shown it resulted in a poor performance of the classifier.

[5]Free translation of the categories is Arts & Style, Career, Economy, Politics, Finance, Mobility, Technology, Opinions, Video

model	input	samples	epochs	learning rate	F1
bert-base-multilingual-uncased	Article text	10k	5	1×10^{-5}	78%
	Article text	10k	5	1×10^{-3}	77%
	Titles only	10k	5	1×10^{-5}	67%
	Article text	**15k**	**5**	1×10^{-5}	**79%**
distilbert-base-uncased	Article text	10k	5	1×10^{-5}	34 %

Table 1: The F1 results of the article category classifier with various configurations of training algorithm.

3.7 Weighted articles

The proposed methodology does not distinguish between articles that truly and in-depth deal with the issue of economic-policy uncertainty and those that only address the issue in a very marginal context. We assume that an article that is entirely devoted to an economic issue should receive more weight than an article that only addresses this problem to a minor extent. Another problem is that it is also possible that a random distribution of words in the text might lead to the fulfilment of conditions. For example, an article that contains many words has a higher chance of satisfying our conditions because a single word from each category may appear in different paragraphs and may also be used in a different context. The likelihood of this error should therefore increase with the inclusion of more words in selected categories (E, P, U).

To partially eliminate this error, that derives from the random occurrence of combinations of words, we suggest the application of a method that attaches weight to the articles. There are various methods of weighting that can be used. We weighted the articles by counting the aggregate number of the selected keywords and if the condition of three combinations is met, i.e., at least one word from each category. Simplistically, if in one month there were 10 articles that met the condition and in the next there were also 10 articles (let's assume that the total number of published articles was constant), the weighting given to these results would be determined by the number of times the specific keywords were repeated in the selected articles in a given month. Formally,

$$EPU_{weighted} = \frac{\sum_{i=0}^{N_t}(e_i + u_i + p_i)}{N_t} \tag{9}$$

where e_i, u_i, p_i represents the number of words in the i-th article that are found in the set of search terms from the domain of economy (uncertainty or policy respectively). N_t is the total number of articles in the observed period t (in our case 1 month).

4 Results

In this section, we will present the results of the robustness tests intended to analyse the potential problems of the newspaper-based methodology proposed by Baker et al. (2016) using data to calculate a German EPU index. The results are presented in comparison with the original approach.

First, in Figure 1, we present the results of tests of how sensitive the EPU index values are to the subjective selection of words from within the three chosen word categories. This is done by gradually adding words to the categories related to economy, policy and uncertainty based on cosine similarity, but without the subjective selection of appropriate words. This will show us how the index responds, either in terms of fluctuations, level, or peaks.

The value assigned to the EPU represents the index calculated with a certain similarity threshold level of ST_{min}. A value of 1 indicates an exact match for the word selection of the EPU authors. Gradually decreasing the value of the threshold level from an EPU of 1.0 to 0.5, represents an expansion in the number of keywords in the individual categories (E, P, U)[6]. For more details, see tables B1–B3 in the Appendix B, where we provide an overview of the individual words in each category that are included at the individual threshold levels.

The results show that the general level of uncertainty, volatility, and major peaks do not fundamentally change and remain stable. For more detail, we present a summary of the statistics and a correlation analysis of the compiled economic policy uncertainty sub-indexes in Tables A1 and A4 in the Appendix A. Overall, the index results vary slightly, with decreasing the level of correlation as more words are included. At the 0.8 threshold, which represents a significant extension to the original query, the index correlates at almost 0.89 in comparison to the original approach. At the 0.5 similarity threshold level, the correlation level with the original index result is 0.71. However, the lower threshold value could result in semantically distant words considered as synonyms. There was no clear association between word expansion and mean index value or standard deviation. Although in all but one case the mean value of the index was higher (0.9 for EPU), and the standard deviation was lower. This suggests that expanding the number of keywords seems to increase the average level of each index, and at the same time, the peaks of the indices are less extreme.

Economic Policy Uncertainty: Application of Different Thresholds

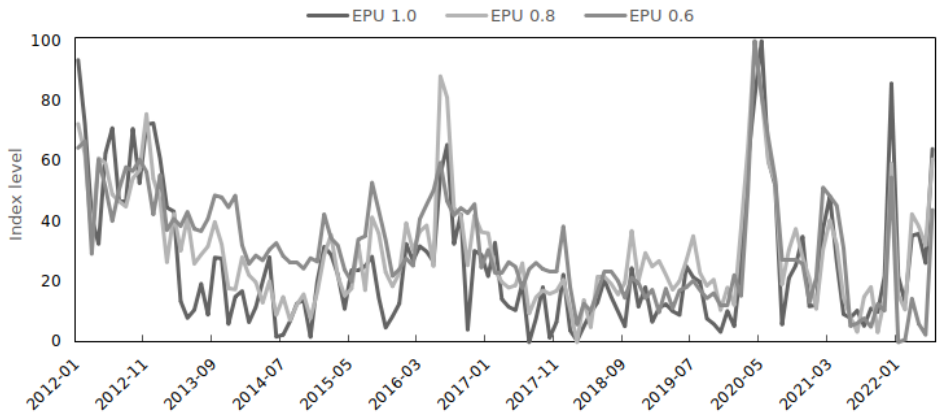

Figure 1: EPU 1.0 is a reproduction of the index as proposed by the original methodology. A value of 1.0 is the threshold level for word selection, where 1.0 equals exact word matches as proposed in the creation of the EPU index. The gradual decrease in the threshold levels values represents an expansion of the number of individual words in the categories (E, P, U). This includes the addition of all words whose cosine similarity was less than a set threshold (including polymorphism and synonyms), that were not included in the original methodology but fit, contextually, into the categories. The index was normalized using min-max feature scaling ranging from 0–100.

In the second exercise, we treat the articles differently with respect to their context, by using the article weighting method presented in section 3.7. We present our results in Figure 2 and further details, the summary statistics, are provided in Table A.2, and the correlations analysis in A.5, within the Appendix A.

We can observe that the level of the original EPU index (EPU) and the magnitude of its peaks is slightly higher in comparison to the weighted approach (EPU weighted articles). But if we com-

[6] In Figure 1, we only present the threshold levels of 1.0, 0.8, and 0.6 for better readability.

pare the index properties as shown in Tables A.2 and A.5 which contain different word queries with respect to cosine similarity (i.e., a decreasing threshold) with the original index statistics in Table A.1, the results show the weighted approach is significantly more stable between selected thresholds and the correlation also remains stronger. This implies that the greater the number of words added to the original methodology, the less stable the results are. We could achieve more stable results by using the method of weighting articles.

When constructing news-based indices, there is no rule that determines the appropriate number of words in each category. However, it does show that the variance of the original EPU index decreases as more words are added while the mean value increases. On one hand, the addition of keywords can improve the identification of articles. On the other, it can also be expected to increase the number of articles that are incorrectly identified in terms of context, or through random condition match. An estimation of the optimal combination of words is not possible using this approach. To deal with this problem, the results show that a weighted approach is more appropriate.

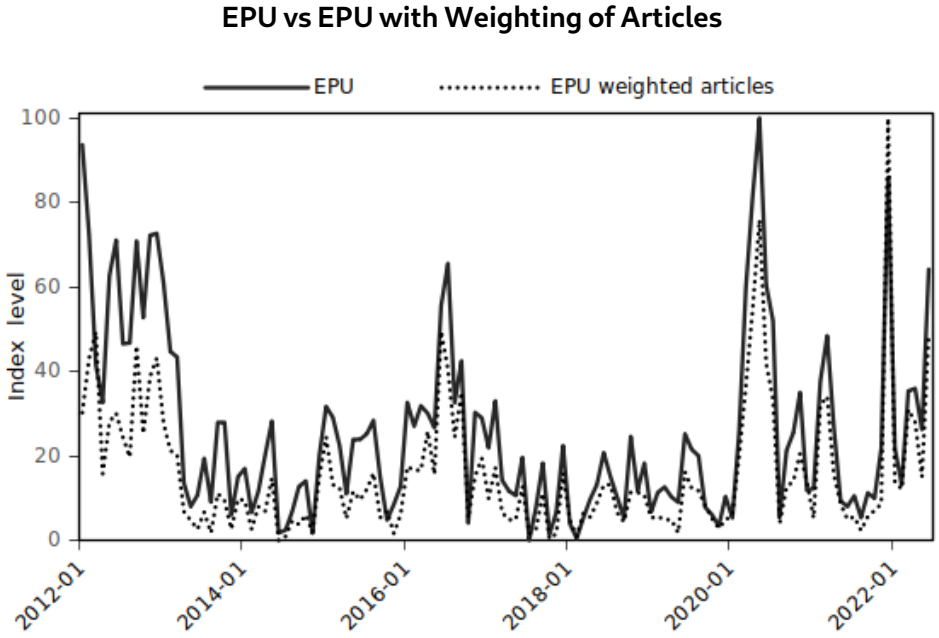

Figure 2: EPU is a reproduction of the index as proposed by the original methodology. The EPU (weighted articles) uses a weighting approach, where the index construction not only takes into account the fulfilment of the condition of containing specific words, but also the number of selected keywords in the article after the condition is fulfilled. The index is normalized using min-max feature scaling ranging from 0–100.

Finally, we present the results of a proposed control for changes in the composition of newspaper articles over time, using only articles that are categorized as economic or political to construct the index, as opposed to the original methodology that considers all categories of articles. The results are shown in Figure 3 and in Tables A.2 and A.3 in the Appendix A. This procedure allowed us to partially eliminate the impact of changes in the content of journals or the preferences of individual editors. Over time, this can affect the ratio of articles fulfilling the keyword conditions to the total number of published articles. This is to be better able to detect a change in trend (in our example, an increase in the level of uncertainty) or a spurious change in trend caused by these changes in preferences.

The results show that the mean value of the index is slightly higher. This is to be expected, as the exclusion of a set of categories is expected to result in a larger change in the numerator-to-nominator ratio. Over longer time periods (e.g., 30 years), EPU indices show a clear upward trend that we do not see in alternative indices constructed as proxies for uncertainty. The data at our disposal does not allow such a long-term comparison using only articles related to economics-politics. Thus, we can observe a change in the properties of the index, but we cannot capture the effect on the long-term trend. These results revealed that after focusing on economic-political articles alone, the indices remain stable with a correlation of over 0.95 for a threshold of 1.0.

EPU index vs EPU Index Compiled with Economic-policy Articles alone

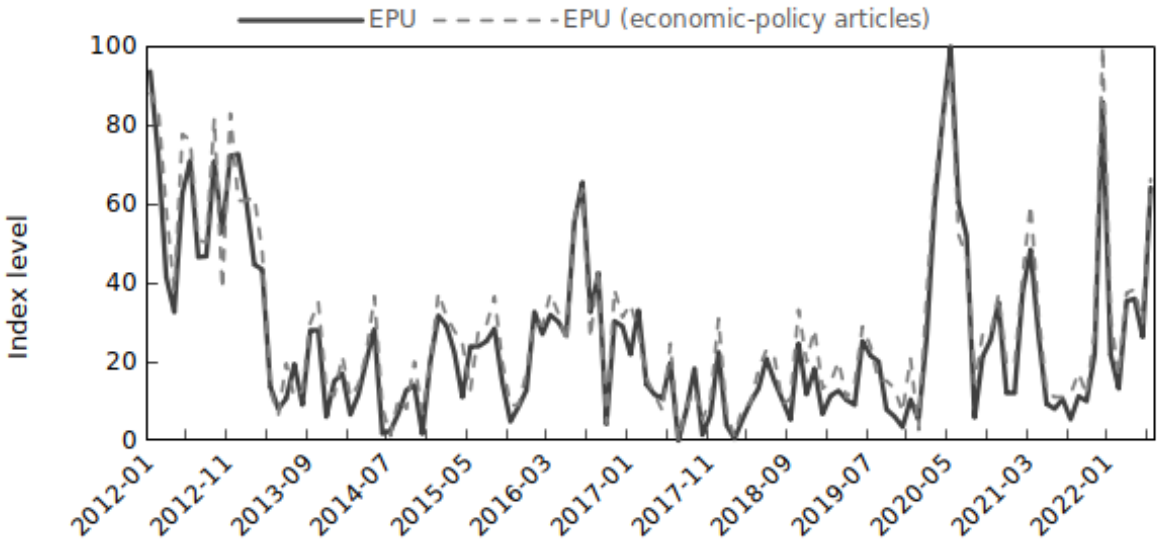

Figure 3: EPU is a reproduction of the index as proposed by the original methodology. The EPU (economic-policy articles) is a construction of the index that focuses on articles classified within the economic-policy category. The index is normalized using min-max feature scaling ranging from 0-100.

Eventually, our data processing allowed us to iterate through all articles of the original German EPU index and provide the distribution of each E, P, and U category in the articles over time. This provides us with information about the relative weight of words in each category over time. Such information was not previously available. We provide greater detail in Figure 4 in the Appendix C.1.

5 Conclusions

Within this paper, we conducted several analyses to test the robustness of the index methodology based on news articles, as per Baker et al. (2016), accounting for the potential problems known to us.

The results show that even though there is no rule for the appropriate number of words in each category (E, P, U) and that the selection of words is highly subjective, the EPU index does not significantly change after word inclusion based on cosine similarity. At a similarity threshold level of 0.8, which is a significant extension to the original query (including polymorphisms and possible synonyms), the index correlated at almost 0.89 in comparison to the original query. However, as the number of words added to each category grows, the mean value of the index

increases slightly and the standard deviation decreases. Thus, expanding the number of words smooths the index. However, if we use a method to weight the individual articles, we can obtain more stable index parameters independently of the number of included words in a search query. In addition to providing a partial solution to the number of words that should be included, the advantage of weighting articles may also reduce the error rate of the original methodology. On one hand, errors may occur through the random distribution of words in the text, which could possibly result in the random fulfilment of a condition for word combinations in articles unrelated to economic-policy uncertainty. The probability of this random match increases with the number of words included in specific word categories. On the other hand, we assume that an article that is entirely devoted to an economic issue should receive more weight than an article that addresses this problem to a very minor extent. The final potential problem we address is the potential for compositional changes in newspaper articles over time. This could impact the ratio of articles that meet the conditions for the selected keywords and the total number of articles published. Although our data series is limited through the use of a start date of 2012, the results showed that changes in the journal composition or in the preferences of individual editors had no significant impact. However, we suggest that this approach could better identify a true change in the overall trend versus a spurious change in the trend caused by these changes in a longer period.

The contribution made by these tests is twofold. Firstly, although these indices are widely used in econometric papers, these potential issues have not yet been sufficiently explored, and it was unknown to what extent these changes may affect the index. Secondly, we believe that the approaches and analyses provided can serve as an approach that will enhance the current newspaper-based methodology.

References

Azqueta-Gavaldon, A., Hirschbühl, D., Onorante, L., and Saiz., L. (2020). Economic policy uncertainty in the euro area: an unsupervised machine learning approach. *SSRN 3516756*.

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105 (3):1177–1216.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2016). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Miranda-Belmonte, H. U., Muñiz-Sánchez, V., and Corona, F. (2023). Word embeddings for topic modeling: An application to the estimation of the economic policy uncertainty index. *Expert Systems with Applications*, 211.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*, 1908.10084.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., and Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International journal of forecasting*, 34(2):355–365.

Vargas-Calderón, V. and Camargo, J. E. (2019). Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. *Cities*, 92:187–196.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2015). *Fundamentals of Predictive Text Mining*. Springer.

Appendices

A EPU summaries

A.1 Statistical summary of EPU and EPU variations

Variable	Obs.	Mean	Std. dev.	Min	Max
EPU 1	130	26.97489	23.03637	0	100
EPU 0.9	130	25.76081	20.76479	0	100
EPU 0.8	130	31.0452	19.21994	0	100
EPU 0.7	130	29.11627	15.55168	0	100
EPU 0.6	130	31.80593	17.45369	0	100
EPU 0.5	130	32.65728	17.92468	0	100

A.2 Statistical summary of EPU, weighted

Variable (w.a)	Obs.	Mean	Std. dev.	Min	Max
EPU (weighted) 1	130	16.66009	16.45487	0	100
EPU (weighted) 0.9	130	19.28014	18.95193	0	100
EPU (weighted) 0.8	130	22.68269	19.26745	0	100
EPU (weighted) 0.7	130	22.03229	17.39956	0	100
EPU (weighted) 0.6	130	24.02101	18.04484	0	100
EPU (weighted) 0.5	130	24.55946	18.01547	0	100

A.3 Statistical summary of EPU eco-policy news

Variable	Obs.	Mean	Std. dev.	Min	Max
EPU (eco-policy) 1	130	30.38006	23.50681	0	100
EPU (eco-policy) 0.9	130	29.07200	21.37257	0	100
EPU (eco-policy) 0.8	130	34.28732	22.57618	0	100
EPU (eco-policy) 0.7	130	32.89777	17.22177	0	100
EPU (eco-policy) 0.6	130	34.43639	18.99696	0	100
EPU (eco-policy) 0.5	130	34.40229	19.73612	0	100

A.4 The EPU index employing different thresholds: correlations

	EPU (1.0)	EPU (0.9)	EPU (0.8)	EPU (0.7)	EPU (0.6)	EPU (0.5)
EPU (1.0)	1					
EPU (0.9)	0.9498	1				
EPU (0.8)	0.8850	0.9470	1			
EPU (0.7)	0.7794	0.8488	0.8611	1		
EPU (0.6)	0.7065	0.7553	0.7689	0.9334	1	
EPU (0.5)	0.7100	0.7563	0.7709	0.9287	0.9978	1

A.5 The EPU index employing different thresholds, weighted: correlations

	EPU 1 (weighted)	EPU 0.9 (weighted)	EPU 0.8 (weighted)	EPU 0.7 (weighted)	EPU 0.6 (weighted)	EPU 0.5 (weighted)
EPU 1 (weighted)	1.0000					
EPU 0.9 (weighted)	0.9700	1.0000				
EPU 0.8 (weighted)	0.9405	0.9798	1.0000			
EPU 0.7 (weighted)	0.8454	0.9054	0.9190	1.0000		
EPU 0.6(weighted)	0.7774	0.8303	0.8459	0.9689	1.0000	
EPU 0.5 (weighted)	0.7725	0.8228	0.8385	0.9644	0.9986	1.0000

A.6 The EPU index eco-policy news correlations

	EPU 1 (eco-policy)	EPU 0.9 (eco-policy)	EPU 0.8 (eco-policy)	EPU 0.7 (eco-policy)	EPU 0.6 (eco-policy)	EPU 0.5 (eco-policy)
EPU 1 (eco-policy)	1.0000					
EPU 0.9 (eco-policy)	0.9516	1.0000				
EPU 0.8 (eco-policy)	0.8802	0.9401	1.0000			
EPU 0.7 (eco-policy)	0.8018	0.8593	0.8638	1.0000		
EPU 0.6 (eco-policy)	0.7145	0.7474	0.7453	0.9139	1.0000	
EPU 0.5 (eco-policy)	0.7225	0.7523	0.7523	0.9102	0.9967	1.0000

A.7 The EPU index employing different sub-indexes: correlations

	EPU 1	EPU 1 (weighted)	EPU 1 (eco-policy)
EPU 1	1		
EPU 1 (weighted)	0.9115	1	
EPU 1 (eco-policy)	0.9717	0.9016	1

B Extended keywords

B.1 Extended set of input search terms for area "Economy"

economy
wirtschaftlich (-); wirtschaft (-); wirtschaftsfreundlich (0.95); wohnwirtschaftlich (0.939); wirtschaftswissenschaftlich (0.937); wirtschaftsbuch (0.935); wirtschaften (0.93); wirtschaftliche (0.927); wirtschaftshilfen (0.925); wirtschaftsfreundlichst (0.922); wirtschaftlichen (0.921); wirtschaftshilfe (0.921); wirtschaftsmagnaten (0.919); wirtschaftraum (0.918); weltwirtschaftlich (0.917); wirtschaftslandschaft (0.917); wirtschaftsgericht (0.917); wirtschaftwoche (0.916); wirtschaftsfeindlich (0.916); wirtschaftmacht (0.915); wirtschaftsaussicht (0.914); wirtschaftshilf (0.914); wirtschaftsmacht (0.914); wirtschaftsmächt (0.914); wirtschaftswissenschaft (0.913); wirtschaftsboß (0.913); wirtschaftsforensisch (0.913); wirtschaftspreß (0.912); wirtschaftlicher (0.912); wirtschaftstreffen (0.912); wirtschaftsries (0.912); wirtschaftliches (0.912); wirtschaftssegen (0.912); wirtschaftsteil (0.911); wirtschaftsgröße (0.911); wirtschaftend (0.91); wirtschaftssystem (0.91); wirtschaftsdaten (0.91); wirtschaftsmarkt (0.908); wirtschaftsweise (0.908); wirtschaftshistorisch (0.908); wirtschaftssachverhalt (0.908); wirtschaftsbereich (0.907); wirtschaftsgut (0.907); wirtschaftstag (0.907); wirtschaftwoche (0.906); wirtschaftsgüter (0.906); wirtschaftswesen (0.906); wirtschaftskrieg (0.906); wirtschaftswelt (0.906); wirtschaftsbericht (0.906); wirtschaftsalltag (0.905); wirtschaftsgut (0.905); wirtschaftsgemeinschaft (0.904); wirtschaftszeit (0.903); wirtschaftsbank (0.903); wirtschaftsnation (0.903); wirtschaftsschwach (0.903); wirtschaftssache (0.903); wirtschaftstreff (0.903); wirtschaftsspion (0.903); wirtschaftszwerg (0.903); wirtschaftsdienst (0.903); wirtschaftsimperium (0.902); wirtschaftssituation (0.902); wirtschaftkrise (0.902); wirtschaftsflaute (0.901); wirtschaftsrisike (0.901); wirtschaften (0.901); wirtschaftsbündniß (0.9); wirtschaftsbuchpreis (0.9); wirtschaftsfachleut (0.9); wirtschaftspakt (0.9); wirtschaftsfreundlicher (0.9); wirtschaftsanalyste (0.9); wirtschaftstag (0.9); wirtschaftskreis (0.899); wirtschaftseliten (0.899); wirtschaftsanwält (0.899); wirtschaftsanwalt (0.899); wirtschafteen (0.899); wirtschaftsmacht (0.899); wirtschaftsmächt (0.899); wirtschaftend (0.898); wirtschaftsszene (0.898); wirtschaftserholung (0.898); wirtschaftsblatt (0.897)

Table 2: The extended (incomplete) set of search terms derived from input words. The value in parentheses represents its similarity to the input (the higher the number, the greater the similarity). A hyphen is provided for the input words.

B.2 Extended set of input search terms for area "Policy"

policy
ausgäben (-); ausgaben (-); ausgaben (-); regulierung (-); steuer (-); defizit (-); zentralbank (-); wirtschaftspolitik (-); haushalt (-); regulierungs (-); ezb (-); EZB (-); haushaltsdefiz (-); bundesbank (-); regulierung (-); wirtschaftspolitike (0.979); regulierung (0.969); regierungs (0.955); zentralbanke (0.944); regierung (0.942); wirtschaftsrisike (0.942); regulierungswut (0.94); regulierungsdichte (0.94); wirtschaftspolitikerin (0.94); regulierungsdichte (0.937); regelung (0.936); regieerneuerung (0.934); wirtschaftsdynamik (0.933); steuer- (0.93); ausgeben (0.93); regierungs (0.929); wirtschaftsproduktion (0.928); regulierungskost (0.927); regulierungsschritte (0.926); regulierungsrisik (0.925); regulierungsseitig (0.924); regulierungsseitig (0.924); regulierungspause (0.923); reglementierung (0.923); regulierungsbürde (0.922); regulierungsregel (0.922); regulierungsbemühung (0.922); regulierungsrunde (0.921); regulierungswut (0.921); regulierungspause (0.92); wirtschaftkrise (0.919); regieerneuerung (0.919); wirtschaftsgenie (0.918); regierung (0.918); wirtschaftstag (0.918); regulierungsschritte (0.918); regulierungskost (0.917); wirtschaftsriese (0.916); wirtschaftskolonne (0.915); regulierungswünsch (0.915); regulierungskiste (0.914); ausgraben (0.914); regulierungsdickicht (0.913); regierungshilfe (0.913); wirtschaftspakt (0.912); wirtschaftspresse (0.912); regulierungsbürde (0.912); wirtschaftnatione (0.912); wirtschaftsdynastie (0.911); regierungsbank (0.911); wirtschaftspreis (0.911); wirtschaftsethik (0.911); haushaltsdefizit (0.911); regulierungsdickicht (0.911); wirtschaftspolitisch (0.911); reglementierung (0.911); wirtschaftswesen (0.911); regelung (0.911); regulierungsbemühung (0.91); regulierungsbedingt (0.91); regulierungsrisik (0.909); regelungsdicht (0.909); wirtschaftsdaten (0.909); wirtschaftsweise (0.908); regulierungspaket (0.908); regulierungskiste (0.907); ausgegraben (0.907); regelungsdicht (0.907); regelleistung (0.907); regierungsbank (0.907); wirtschaftsleistung (0.906)

B.3 Extended set of input search terms for area "Uncertainty"

uncertainty
unsicher (-); unsicherheit (-); unsicherer (0.898); unsicheren (0.886)

C Keywords appearance

C.1 Number of keywords in EPU categories

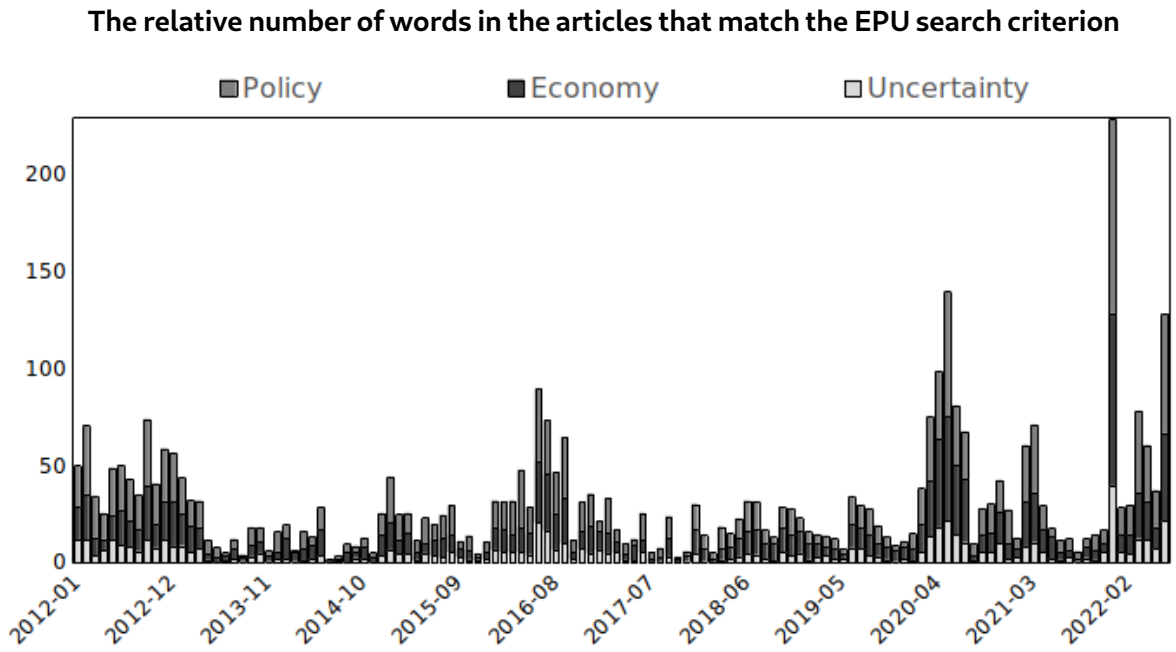

Figure 4: Presented index represent reproduction of EPU index proposed by the original authors. We iterated through all the articles where condition of keyword combinations is met, and counted the number of occurrences of the search terms used in the groups E, P and U. The index was normalized using min-max feature scaling ranging from 0-100 for each category.