Gillmann, Niels; Kim, Alisa

Conference Paper

Quantification of Economic Uncertainty: a deep learning approach

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Gillmann, Niels; Kim, Alisa (2021) : Quantification of Economic Uncertainty: a deep learning approach, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/242421

Quantification of Economic Uncertainty: a deep learning application

Alisa Kim[a], Niels Gillmann[b,c,*]

[a]*School of Business and Economics, Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany*
[b]*Dresden branch, ifo Institute, Einsteinstr. 3, 01069 Dresden, Germany*
[c]*Faculty of Economics and Business, TU Dresden, Münchner Platz 3, 01062 Dresden, Germany*

Abstract

Research on the measurement of uncertainty has a long tradition. Recently, the creation of the economic policy uncertainty index sparked a new wave of research on this topic. The index is based on major American newspapers with the use of manual labeling and counting of specific keywords. Several attempts of automating this procedure have been undertaken since, using Support Vector Machine and LDA analysis. The current paper takes these efforts one step further and offers an algorithm based on natural language processing and deep learning techniques for the quantification of economic policy uncertainty. The new approach allows an accurate distillation of the latent "uncertainty" underlying newspaper articles, enables an automated construction of a new index for the measurement of economic policy uncertainty, and improves on existing methods. The potential use of our new index extends to the areas of political uncertainty management, business cycle analysis, financial forecasting, and potentially, derivative pricing.

Keywords: Economic Policy Uncertainty, Deep Learning, Natural Language Processing, Text Data, Forecasting

1. Introduction

Leading economic experts agree that the slowdown of world economic growth during the period 2018-2019 can be mainly attributed to high uncertainty about political decisions (Tripier, 2019). Events like the "trade war" between the US and China, Brexit, US sanctions against Iran, and the demonstrations in Hong Kong fall into this period. One of the main channels through which the high level of uncertainty affects world economic growth is a falling investment rate of private companies around the world (Bobasu et al., 2020). Economic researchers are still debating about the exact effects of uncertainty shocks on economic activity. Economic theory puts forward precautionary savings as the most prominent explanation (Kimball, 1990). This theory states that when uncertainty increases, actors will put their activities on hold until there is more clarity (Leduc and Liu, 2016). Contrarily, the risk premium theory mentioned

We would like to thank seminar participants at HU Berlin, NUS and TU Dresden as well as conference participants at the Virtual ISF 2020 for their helpful and constructive feedback.

*Corresponding Author
Email addresses: kolesnal@hu-berlin.de (Alisa Kim), gillmann@ifo.de (Niels Gillmann)

in Christiano et al. (2014) argues that the effect of increased uncertainty can even be positive in specific scenarios. Bloom (2009) suggested a "wait-and-see" effect. However, at the moment, no dominating theory can be established.

One of the main reasons why researchers cannot agree on the effects of uncertainty shocks is that there is disagreement on how to empirically measure uncertainty. Recent literature proposes several proxies: volatility of the stock market (Bloom, 2009), dispersion in forecasts of professional forecasters (Glas, 2019; Liu and Sheng, 2019; Sill, 2014), disagreement in the expectations of survey participants (Bachmann et al., 2013; Claveria, 2019), as well as some data-driven approaches (Jurado et al., 2015). The first three proxies share the same shortcoming: they measure perceptions of individual uncertainty instead of the general underlying uncertainty in the economy. This personal perception tends to differ from the aggregated uncertainty in the economy, especially during periods of high volatility, when the formulation of expectations about the future is nontrivial. The latter proxy is trying to overcome this problem by aggregating individual information. Its potential shortcomings lie in the large amounts of economic data required, which may cause slower response to the change in underlying uncertainty since most data have a publication lag of at least one or more months.

Most recently, economists adopted text data as a source to obtain additional information about the economy. The big advantage of this data source is that it is available in a timely fashion and contains much information. Alexopoulos and Cohen (2015) and Baker et al. (2016) presented the first papers that used text to quantify economic uncertainty. The former one uses the new york times newspaper to construct an index of uncertainty in the economy. The latter one improves on the first paper and builds an economic policy uncertainty (EPU) index using ten newspapers. The EPU index is based on the share of articles classified as uncertain in the pool of general articles per newspaper per month. To find the share of articles about uncertainty, the authors had to come up with a set of rules to identify an EPU article. They used manual labeling to create a dictionary that allowed easy construction of their index.

The index by Baker, Bloom and Davies (2016) (BBD) has become widely accepted as a reasonable proxy of uncertainty. But since their publication, methods for using text as data have drastically improved and manual labelling is no longer the only feasible way to construct an EPU index from text data. Since 2017, advanced text mining methods have been applied in the economic literature. One popular method is the Latent Dirichlet Allocation (LDA), an unsupervised algorithm that facilitates identifying latent topics in a document without pre-labeling the data and therefore removing the arbitrary choice of keywords. After identifying the topics in a set of newspaper articles, the researcher can choose those she considers relevant and construct an index from them. Examples are Azqueta-Gavaldón (2017), Larsen (2017), and Thorsrud (2018). Unfortunately, the topics resulting from LDA are not automatically labelled and do not necessarily match the theoretically desired topics. This improves the problem from arbitrarily defined keywords to arbitrarily defined topics but does not remove the arbitrary choice completely. Supervised learning provides a way of identifying relevant keywords inside the text corpus without manual definition and arbitrarily constructed topics. For example, Tobback et al. (2018) used Support Vector Machine (SVM) and applied it to the corpus of six Belgian newspapers over the time from 2000 until 2013. They restricted their initial sample of newspaper articles to those talking about uncertainty in Belgium or the EU. The SVM-based classification model was used to predict

the binary label (containing or not containing economic policy uncertainty) of every article and reconstruct the index using the BBD methodology. The resulting time-series had superior predictive power over some of the Belgian macro indicators like bond yield, the credit default swap spread and consumer confidence as opposed to a Belgian EPU index constructed by keywords as in the original BBD method.

Our goal is to improve the measurement of EPU, removing the need for arbitrarily chosen keywords completely, removing the need for costly manual labeling of keywords through labeling by machine learning and thereby making the index extendable into the future and less costly to construct. In pursuing this goal, the paper contributes to the literature on measurement of economic uncertainty and on the literature of natural language processing (NLP). We introduce a state-of-the-art deep learning model for textual analysis to label newspaper articles according to uncertainty, show that the language in newspaper articles about EPU changes over time and provide evidence that a deep learning method to classify newspaper articles might improve the forecasting ability of the existing EPU index.

In line with the announced goals, we have formulated three research questions (RQ) that define the empirical design:

- RQ 1: can a deep learning classifier learn to label EPU articles without using any keywords but the textual semantics of a newspaper article instead?

We have considered the recently developed NLP models that make use of deep learning (DL) and transfer learning (Radford et al., 2019). The proposed binary classifier distinguishes between articles containing or not containing EPU. We train the model on an article corpus labeled according to the BBD methodology, given its wide adoption by practitioners (Ghirelli et al., 2019; Soric and Lolic, 2017; Zalla, 2017), its carefully conducted manual audit in the time from 1985 to 2012 and the absence of non-manual labeling alternatives. We compare the performance of the proposed approach with some well-known algorithms like SVM and Random Forest, as well as test its robustness with 10-fold cross-validation with stratification (given a major target label imbalance).

- RQ 2: how does the language inside the EPU articles change over time?

To evaluate the temporal dynamics of the newspaper vocabulary, underlying the concept of uncertainty in analyzed corpora, we analyze which words of the input article were considered the most important by the classifier. We perform this task for 1000 EPU articles from every year, select the top ten words per article, assign them a rank from 10 to 1 and then sum up the ranks of the entire vocabulary. We further select the ten highest ranking words that will represent the ”uncertainty drivers” for the analyzed year.

- RQ 3: can the selected DL methodology show better adaptability to the changing rhetoric than the BBD index?

To explore the adaptive capacities of the model, we have transformed the values predicted during cross-validation into an index using the original BBD methodology. We obtain two indices: one reconstructed from the values predicted by our NLP method and one reconstructed the keywords suggest by BBD. We offer a comparative analysis of the two

time-series indices. Firstly, we explore their co-movement with alternative uncertainty proxies to identify suitable benchmarks, and then assess their predictive power over a set of key macroeconomic variables.

2. New deep learning-based EPU index

Our method is based on the approach by BBD which consists of a manual audit of newspaper articles to identify relevant keywords and then using these keywords on newspaper data to construct an EPU index. The choice of keywords is arbitrary and we try to overcome this choice of keywords by applying deep learning to the method of BBD. In this section, we describe the original methodology to construct an EPU index in more detail, then discuss our data and explain the reconstruction procedure of a new DL-based EPU index.

2.1. Original BBD methodology

Baker et al. (2016) argue that the share of newspaper articles about EPU in every month should be a good indicator for uncertainty. To construct this indicator, they use a large database of ten newspapers in the United States. In this database, they count the number of articles about EPU and about the economy in general for every month, obtaining a set of EPU articles and a set of economic articles. The counting is based on the presence of three sets of keywords: ”economy or economic” + ”uncertainty or uncertain” + a term from a group referring to policy (Congress, deficit, Federal Reserve, legislation, regulation, White House). If an article contains at least one keyword from all three groups, it is labeled as containing EPU (target=”1”). The share of EPU articles when applying the BBD keywords to our data can be seen in Figure 1:

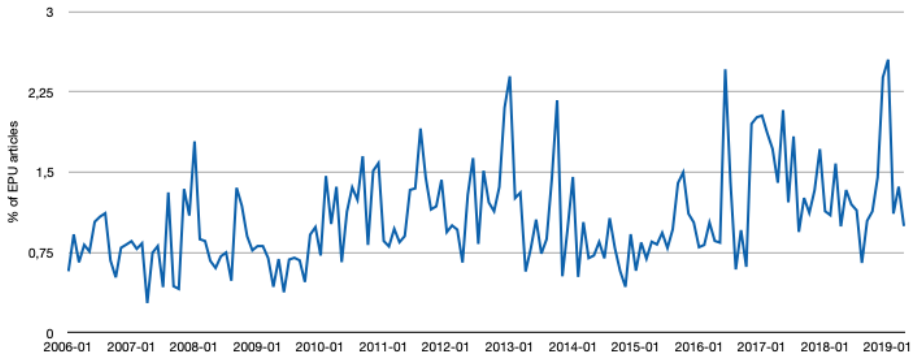

Figure 1: Monthly share of EPU articles in the Sample.

The set of keywords that BBD use for labeling was derived from an extensive manual audit of a large corpus of articles from five leading US newspapers during the time from 1985 to 2012. Since the use of the ”economic” and ”uncertainty” keywords is undisputed, the BBD authors decided to limit their sample for the audit study to articles containing the terms for ”economic” and ”uncertainty”. The audit study then helped to narrow down which policy terms are necessary to identify relevant EPU articles. In detail, groups of researchers were reading random samples of

newspaper articles from the collection and labeling them as either related to EPU, not related to EPU or hard to tell. BBD refined the set of keywords until both human labeling and search requests yielded the same EPU index.

The output of the labeling is a binary classification of newspaper articles into those containing economic policy uncertainty (target=1) and those with general economic news (target=0). This classification can be transformed into a monthly index by calculating the share of EPU articles per newspaper per month and standardizing the share for each newspaper individually. Additionally, the average over all newspapers for each month is taken and then standardized to a mean of 100.

The main drawback of the method proposed by BBD is the set of fixed keywords. While it is easy to implement, it inevitably leads to an oversimplification and a somewhat arbitrary choice of keywords. BBD do their best to reduce the arbitrariness of their keyword choice with the manual audit study. Outside of their manual audit period (1985-2012) however, we cannot be sure that their suggested keywords actually pick up on all the articles they should. This can occur if for example the vocabulary that newspaper use changes over time and one of the keywords is not used anymore even though the article still covers the topic of EPU.

2.2. An Index based on deep learning

We are trying to overcome the arbitrary choice of keywords by implementing a deep learning algorithm for identification of EPU articles. Therefore, we need a much larger set of full-text articles than BBD needed for their manual audit. Because of the need for full-text articles and their limited availability, we had to use a different set of newspapers than BBD. The use of all articles in our sample period did not appear possible due to the large volumes. Thus, we had to limit our sample to include only relevant articles. We followed the example of Tobback et al. (2018), assuming that all relevant articles must contain the ”economy” or ”economic” keywords. Baker et al. (2016) performed this preselection of newspaper articles for their audit study to an even greater extent, only including articles which contained both the economic and uncertainty keywords and collected 12 009 full-text articles. The most recent contribution in the literature by Tobback et al. (2018)) offers an analysis of 210 000 full-text articles. Our analysis is done on 315 543 articles, from 01 Jan 2006 to 30 Apr 2019, offering the biggest text corpus for labelling newspaper articles according to EPU so far. The start date before the Global Financial Crisis (GFC) is selected in order to capture both periods with normal and high levels of EPU. We aspired to include newspapers that guarantee coverage across the whole of the USA. Our articles come from The Washington Post, Pittsburgh Post-Gazette (Pennsylvania), The Atlanta Journal-Constitution, St. Louis Post-Dispatch (Missouri), The Philadelphia Inquirer (Pennsylvania), USA Today, Star Tribune (Minneapolis, MN), The Orange County Register (California), Tampa Bay Times (Florida, previously known as St. Petersburg Times) and The New York Post. The distribution of articles across newspapers is shown in Table 1:

The uneven distribution of articles is partially explained by the size of newspapers' editorial offices and a large amount of reprints and reposts of existing articles (NYP in particular) that were dropped from the sample. The articles are used for modeling without regard to the source, while the index reconstruction method accounts for the distribution skews.

Table 1

Number of economic articles per newspaper, 01 Jan 2006 - 30 April 2019.

Newspaper	Number of articles
The Washington Post	81 734
Pittsburgh Post-Gazette (Pennsylvania)	41 225
Tampa Bay Times	36 436
USA Today	26 267
The Atlanta Journal-Constitution	26 038
St. Louis Post-Dispatch (Missouri)	25 400
The Philadelphia Inquirer (Pennsylvania)	22 502
Star Tribune (Minneapolis, MN)	21 422
The Orange County Register (California)	19 983
The New York Post	14 536

To construct our new index, we are employing a DL algorithm to predict the label of newspaper articles as EPU or non-EPU. The initial labeling of the train and test sets was performed according to the keywords established by BBD. The final labelling obtained by the DL algorithm was aggregated to an index time-series in the same way as in the BBD paper.

3. Methodology

DL applications for textual data in economics are yet sparse. Thus, we will revisit the principles of DL and NLP, as well as provide a detailed configuration of the selected classification model. Following RQ 1, we are offering a classification model that is capable of identifying uncertainty in newspaper articles without a fixed set of keywords. This model is based on DL and NLP techniques, namely a recurrent neural network that uses GPT-2 pre-trained embeddings (Radford et al., 2019) and an attention mechanism (Vaswani et al., 2017). This section begins with the description of the data pre-processing steps, followed by the introduction of the embeddings concept and the GPT-2 language model. We further provide clarifications on the DL architecture and elaborate on the attention mechanism. The latter will be instrumental for RQ 2, when we address the change of the ”uncertainty drivers” over time.

3.1. Data pre-processing

In order for an NLP model to process text, the words are converted into a numeric representation. We analyzed the average lengths of the article body and the headline. Table 2 shows that EPU articles tend to be longer. This particularity is accounted for during the text pre-processing.

Table 2

Average number of words in E- and EPU-labeled articles and corresponding headlines before preprocessing.

	Average length of the article body	Average length of the headline
All	823	9.31
E articles	817	9.2
EPU articles	1 087	10.2

The corpus vocabulary has to be carefully considered in order to facilitate the task of knowledge extraction. This entails homogenizing and cleaning the provided textual data from noise. The headlines were integrated into the text. The pre-processing steps included three main stages. The first stage comprised of vocabulary filtering: opening up and converting the contraptions (”can't” into ”cannot”) and removal of the usual stopwords (excluding negations (”not”)). All words that occur less than ten times were also dropped (bringing the vocabulary size from 468 997 to 114 763), which allowed accounting for misspelling as well. Importantly, train and validation sets were stripped of the keywords ”policy or political”+”uncertainty or uncertain” to ensure that the classifier does not learn only based on their presence or absence. During the second stage, numbers and irrelevant components like internet links and punctuation were removed. During the third stage, text got transformed to lower case. As a result of pre-processing, the average length of the article shrank to 407 words. The cleaned article text is broken into a list of words (tokens).

3.2. Natural Language Processing: language models

NLP focuses on the methods that allow machines to analyze and evaluate human language. The task of text representation in a numeric format lies at the basis of NLP. However, modeling a system as complex and intricate as a human language proves to be a very complex task, even with the appearance of large digital corpora of text in the 90s. Teaching computers to understand the written text involved the necessity of approximating the irregular structure of the human expression and modeling language rules, leading to the introduction of Language Models. Nowadays, Language Models are used in machine translation, text classification, speech recognition, handwriting recognition, information retrieval, and many other (Bahdanau et al., 2014; Graves et al., 2013; Hirschberg and Manning, 2015).

Two main classes are statistical and neural Language Models. The first class uses traditional statistical techniques like N-grams and linguistic rules to learn the probability distribution of words in a studied text (one of the early examples is Bahl et al. (1989)). Widely used solutions included one-hot encoded bag-of-words vector representations and the TF-IDF representations (also known as frequency embeddings, Salton et al. (1975). The latter represents the matrix of document vectors, containing term occurrence frequencies (TF) or their transformation by weighting with the inverse document frequency (IDF). The key idea of TF-IDF representation lies in the assignment of larger weights to words with higher discriminatory ability. This principle entails that frequent occurrence of a term in the document does not lead to high importance; rather, the word must be unique for that document at the corpus level. This ranking is widely used in NLP, in particular, for sorting data into categories, as well as keywords extraction.

The second class became a new powerful tool for NLP with the adoption of neural networks to model language (Bengio et al., 2003). This area saw tremendous developments in recent years and became industrial state-of-the-art, used in Google translate, virtual assistants like Apple's Siri and Amazon's Alexa. We have applied the solutions developed by OpenAI, who released a new language model called GPT-2 in 2019. GPT-2 is a transformer-based generative language model that was trained on 40 GB of curated text from the internet (Radford et al., 2019).

3.3. Embeddings

In order to preserve the semantic meaning and linguistic characteristics of a word, we can transform it into a vector representation, called word embedding. Although known before (frequency-based embeddings and vector-space model), the concept of word embeddings re-emerged in 2013 with the introduction of prediction-based "neural" embeddings. The work of Mikolov et al. (2013) started a new chapter in the development of the field, allowing to represent words as numeric vectors without the sparsity of one-hot encoded matrices and retention of the semantic meaning as opposed to TF-IDF representations. The proposed word2vec is an advanced model for word embedding, composed of a neural network model that is capable of learning word representations during training on a large text corpus. Mikolov et al. (2013) offer two types of training task for the procurement of embeddings: CBOW and Skip-gram. The former forces the model to predict a target word from a window of adjacent context words, while the latter entails prediction of a context window from the provided target word. The resulting word vectors ("inflation"=[0.5, -0.0123, ... 2.1]) are located within the multi-dimensional vector space in such a way that words sharing common contexts within the corpus are positioned next to each other.

The initial word2vec algorithm was followed by GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2016), and GPT-2 (Radford et al., 2019), as well as the appearance of publicly available sets of pre-trained embeddings that were acquired by applying the above-mentioned algorithms on large text corpora.

The GPT-2 pre-trained embeddings used for the proposed model, were trained on 250 thousand documents from the WebText, as stated by Radford et al. (2019). A machine learning method, where a model developed for a specific task, is reused and becomes a starting point for a model on a different task, got known as transfer learning. As defined by Goodfellow et al. (2016), "transfer learning and domain adaptation refer to the situation where what has been learned in one setting ... is exploited to improve generalization in another setting". Usage of pre-trained embeddings proved to be useful for achieving a superior performance in most NLP tasks (Dai and Le, 2015; Peters et al., 2018; Radford et al., 2018; Howard and Ruder, 2018). Given the limited size of our newspaper corpus, we use word embeddings that were trained on a much larger sample as part of our model for EPU classification. To that end, we replace the words of an article with its pre-trained embedding feature vector. This approach maintains the word order in an article. Given the sequential nature of the data, the architecture of a classifier plays a critical role in obtaining a prediction accuracy.

3.4. Deep Learning: recurrent and bidirectional neural networks

DL is a subset of Machine Learning primarily based on the hierarchical approach, where each step converts information from the previous step into more complex representations of the data (Goodfellow et al., 2016). We refer to deep learning when the used Artificial Neural Network uses multiple layers (Deng and Yu, 2014). DL methodology aims at learning multiple levels of representations from data, with higher levels reflecting more abstract concepts, thus capturing the complex relations between the data set features (Kim et al., 2020). This ability made DL a popular solution for a wide range of modeling tasks. The adoption of DL methods in scientific areas like economics, however, was limited by the necessary computational capacities and interpretability issues. Neural networks notoriously represent a 'black box' - a shortcoming originating of its inherent internal complexity (Gilpin et al., 2018).

Regardless of these shortcomings, the development of DL offered a versatile toolbox for the processing of sequential data i.e., time series and text. New DL architectures like convolutional neural networks (Kalchbrenner et al. (2014)), recurrent neural networks (RNN, Hochreiter and Schmidhuber (1997)), Hierarchical Attention Networks (Yang et al. (2016)) were successfully employed for learning textual representations (Krauss et al., 2017). In particular, Athiwaratkun and Stokes (2017); Yin et al. (2017); Zhang et al. (2018) showcase the ability of RNN variation like Gated Recurrent Unit (GRU) to show improved performance on NLP tasks. As opposed to RNN that may fail to capture the long-term information due to the gradient vanishing problem, the GRU is equipped with a set of "gates" that allow GRU to dynamically remember and forget the information flow, which is crucial for longer text inputs (Cho et al., 2014).

Addressing the non-linear nature of text understanding, Schuster and Paliwal (1997) suggested a further reinforcement of the RNN with a bidirectional component. For the case of uncertainty classification, analyzing the preceding, as well as the following observations, is equally important for the extraction of the semantic concepts (Liu et al., 2020). Their ability to grasp long-term dependencies motivated the choice of a bidirectional GRU layer as a significant component of the suggested classifier. The full DL architecture is represented in Figure 2, where x_1 to x_T represent the textual input transformed into *tokens*. As mentioned before, the average length of an article is 407 words, which was established as a fixed input length (T=407). In order to feed in the article into the DL model, text strings must be numeric. We transform every word into a token ("inflation" $->$ "34") and create a lookup vocabulary that allows to map the tokens back to words. We further truncate longer articles and *pad* shorter articles. Padding means adding fixed values (in our case "0", which doesn"t have any semantic meaning to it) in the beginning of an article until it reaches a length of 407 tokens. Truncating means removing the words which are too much from the bottom of each article. This is justified since newspaper articles usually have the most relevant information included in the top and not in the bottom of the article.

The output is represented by the single neuron with sigmoid activation, given the binary classification task. The model outputs probabilities for the supplied array of *tokens* representing an article to be containing EPU (target=1). The layer that follows the input is a dense matrix of embeddings. As discussed above, we use a set of pre-trained GPT-2 embeddings. Every word in the dictionary (114 763 words) is assigned an embedding vector of 768 neurons

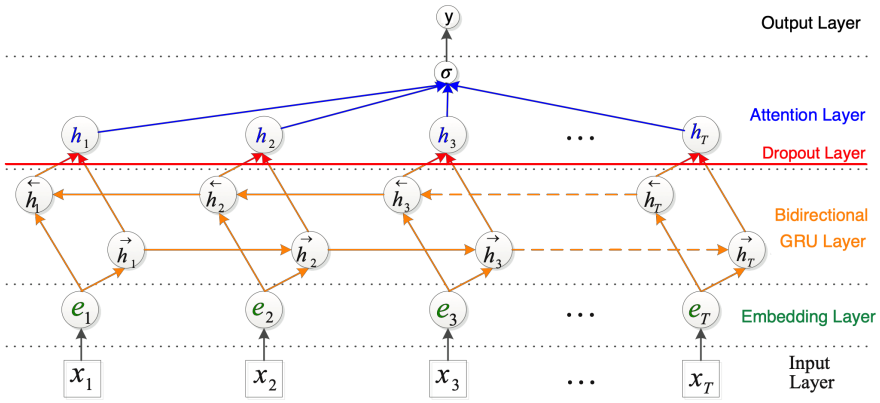

Figure 2: Architecture of the proposed DL model for EPU classification (based on illustration provided by Zhou et al. (2016))

(defined by the authors of the GPT-2 language model). Thus, the embedding matrix has dimensions 768 x 114 763 and functions as a look-up table. Input integers are used as the index to access this table. We have e_t vectors as output, each representing an input word. These vectors are supplied to the bidirectional GRU layer that will process them word by word. This layer's output will be a hidden state h^t vector that will go into the dropout layer (also depicted in Figure 2) and further passed to the attention layer.

Equations 1-4 showcase the internal functionality of a GRU layer. As opposed to LSTM, GRU does not have a component called *cell state* and uses the *hidden state* to transfer information (Cho et al., 2014). It also has only two gates: a *reset gate* and *update gate*. The reset gate is used to decide how much past information to forget, while the update gate is used to decide which information will be discarded or added. Equation 1 defines the reset gate, Equation 2 - the update gate, and Equations 3 and 4 describe the transformations to obtain the hidden state. The single-layer GRU computes the hidden state h^t for word x^t with W and U representing weight matrices and b bias vectors of corresponding elements of the GRU cell, $\odot$ denotes the element-wise multiplication of two vectors:

$$r^t = \sigma(W_r x^t + U_r h^{t-1} + b_r) \tag{1}$$

$$z^t = \sigma(W_z x^t + U_z h^{t-1} + b_z) \tag{2}$$

$$\tilde{h}^t = tanh(W_h x^t + U_h(r^t \odot h^{t-1}) + b_h) \tag{3}$$

$$h^t = z^t \odot h^{t-1} + (1 - z^t) \odot \tilde{h}^t \tag{4}$$

As we are using a bidirectional GRU, the network will contain two sub-networks for the left and right sequence context, which develop forward and backward, respectively. The output of the t word is thus represented by the

element-wise sum that combines the forward and backward pass outputs:

$$h^t = \overrightarrow{h_t} \odot \overleftarrow{h_t} \tag{5}$$

The hidden states from the bidirectional GRU layer will be further passed on to the dropout and attention layers. The output of the attention layer s is supplied into the output layer with a sigmoid activation, that produces the probability of the article a, containing T words, to be containing EPU:

$$y_a = \sigma(W_o s + b_o) \tag{6}$$

The binary cross-entropy loss is used for end-to-end training:

$$\mathcal{L}(y_a, \hat{y_a}) = -\frac{1}{T}\sum_{i=1}^{T}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log(p_{ij}) \tag{7}$$

3.5. Regularization measures: dropout layer

Like all complex systems, neural networks are vulnerable to overfitting (Hinton et al., 2012). To make sure that the model learns to generalize from the training set without picking up the noise, our GPT-2 deep neural network (DNN) includes a dropout layer after the bidirectional GRU layer. The concept of dropout comprises removal at random of hidden layer neurons and their corresponding connection weights during training. The probability of a hidden neuron being dropped out follows a Bernoulli distribution with a given dropout rate, in our case, a 50% chance.

3.6. Attention layer

Another important component of the proposed model from Figure 2 is the attention layer (Vaswani et al., 2017). As pointed out by Zhou et al. (2016), attention has been successfully adopted for several NLP-related tasks, like reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations. An attention function is mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are words vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

We are using a weighted average attention mechanism as applied by Chorowski et al. (2014) and Zhou et al. (2016), which produces a weight vector and merges word-level features from each time step into a sentence-level feature vector, by multiplying the weight vector. The calculation is depicted in Equations 8 and 9: let H be a matrix consisting of output vectors [h_1,h_2,...,h_T] from the bidirectional GRU layer and w - a trained parameter vector after Dropout was applied. T remains a sentence length of 407 words. The weighted sum of these output vectors forms the representation s of the sentence:

$$\alpha = softmax(w^T H) \tag{8}$$

$$s = H\alpha^T \tag{9}$$

An additional value of the attention layer stems from its interpretation features that will be explored further.

4. Results

In this section, we first evaluate our models for classification of EPU articles. All data-processing and modeling computations are performed in `python` with the use of packages like `numpy`, `pandas`, `scikit-learn`, `nltk`, `gensim`, for DL implementation the high-level neural network library `keras` is used as well as the `transformers` package by HuggingFace. Then, we show the improved information content of our new index compared to the BBD keywords-based uncertainty index by looking at correlation with key macroeconomic variables and a forecasting exercise.

4.1. Classification analysis

According to the experimental design, we have developed a DL-NLP-model that allows the accurate classification of articles according to the previously discussed labeling. The test set represents 30% of the data and contains approximately 2% of EPU cases, matching the label balance of the train set. For evaluation, we have selected the AUC (Area under the Curve) and the F1-score. The former reflects how much a model is capable of distinguishing between classes regardless of the threshold. and is robust toward class imbalance. The latter allows evaluating the accuracy of the predictor by considering both precision (number of correct positive results divided by the number of all positive results) and recall (or sensitivity, correct positive results divided by the number of all relevant samples) of the test set. The F1-score represents a harmonic mean and measures how precise and how robust the models classify EPU cases:

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

We considered a variety of different neural network architectures for the model training process. The neural networks with bidirectional and attention layers provided the best performance in the selected metrics. Apart from GPT-2, we have tried two other widely-used pre-trained embeddings: Google News Embeddings and GloVe. The results were inferior to GPT-2 and will not be discussed further.

Table 3 shows the results of the GPT-2 DNN model and selected benchmarks: TF-IDF vector-based logistic regression (LR), SVM, Random Forest (RF) and XGBoost (XGB). GPT-2 DNN outperforms other models with the highest AUC of 0.96, but its improvement for the F1-score is even more substantial, reaching 0.65 as compared to other models. Tree-based models seem to be particularly weak with the precision and recall. LR and SVM with non-linear kernel capture the case of interest more accurately. Given that the classifiers were trained on the dataset without the original keywords, and considering the strong performance of GPT-2 DNN, Table 3 allows us to conclude that RQ 1 was answered positively: we have successfully constructed a DL model that can capture the concept of EPU using text mining. However, to examine the robustness of the proposed solution, we performed 10-fold stratified cross-validation. Table 4 shows that GPT-2 DNN keeps up the excelling performance with an AUC standard deviation of 0.014 and an F1-score standard deviation of 0.04. However, the heterogeneity of input is visible through the folds, regardless of the randomized splitting.

Table 3

Evaluation of classifier models on the randomized out-of-sample test set.

Models	AUC	F1-score
LR	0.9116	0.1550
SVM	0.8966	0.2083
RF	0.9063	0.0356
XGB	0.9054	0.0171
GPT-2 DNN	**0.9606**	**0.6500**

Training of the benchmarks was performed with `scikit-learn` package and the following hyperparameters: LR - L2 regularization and the "lbfgs" solver; SVM - regularization parameter C =1.0, "rbf" kernel, gamma=0.0024; RF - number of estimators=100, maximum depth=None, minimum samples split=2, maximum features=20; XGB - learning rate=0.1, number of estimators=100, maximum depth=3.

Table 4

Results of the 10-fold cross-validation (with stratification of samples) of the GPT-2 DNN.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	**Mean**
AUC	0.9507	0.9725	0.9553	0.9587	0.9316	0.9760	0.9606	0.9557	0.9722	0.9829	**0.9616**
F1-score	0.7281	0.7068	0.6500	0.6294	0.5787	0.6613	0.6571	0.6567	0.6584	0.6698	**0.6590**

To further examine the potential presence of narrative shifts (topics, used vocabulary) in EPU articles over time, we have looked into the decision-making mechanism of GPT-2 DNN, in particular, its attention layer. The next section illustrates the analysis of the changing semantics in the newspaper articles over time.

4.2. Evolution of uncertainty rhetoric

RQ 2 concerned the potential shortcomings of a static keyword approach. Our goal was to analyze if there is a change in the words that entail EPU. We have sub-sampled all the EPU articles by year (on average, 235 articles per year), dropped the three groups of EPU keywords, and used them as a test set for the trained classifier. We extracted the weights assigned by the attention layer to the word inputs after the model was trained i.e., in the inference phase. The top ten words with the highest values were selected and assigned points from ten to 1. Points accumulated during the year constituted a ranking of every word by its "uncertainty impact". The top ten words for every year are showcased in Table 5. One can observe an evident change of the newspaper agenda and the introduction of new "uncertainty drivers" through time. The years 2008-2009 are focused on economic "crisis" and "recession", followed by concerns on "fiscal" policies and changes in "legislation". The pre-election years see the rise of national agenda with "american" and "america" leading and "trump" first appearing in 2015 and firmly dominating the ranks from 2016 to 2019. In 2016 "brexit" enters the ranks, followed by "tariffs" and "immigration" in 2018. Change of "uncertainty drivers" in time indicates a strong interpretation capacity of the DNN classifier and demonstrates its ability to adapt to the new topics with time. Further, the presented evidence raises concern if a set of fixed keywords is enough to capture uncertainty during different periods like the financial crisis in 2009, the trade-war in 2018, or the COVID-19 pandemic in 2020.

4.3. Adaptability analysis

In this section, we compare our reconstructed index to the proxies established in the literature and identify meaningful benchmarks. Our goal is to assess the ability of our index to explain variation in real macroeconomic variables.

There are several approaches to building an uncertainty proxy. They can be assigned into four different categories: proxies based on the number of search requests or newspaper articles during a certain period (Baker et al., 2016), proxies based on variation in a large group of macro variables (Jurado et al., 2015), proxies based on disagreement in expectations among survey participants (Ozturk and Sheng, 2018), and proxies based on the volatility of economic variables (Bloom, 2009).

BBD represents the first category. The survey-based uncertainty index relies on data from the consensus survey, an aggregator that collects surveys of economic forecasters from many different sources. The forecast error of the different survey participants can be interpreted as a proxy for uncertainty in the economy. The macro-based uncertainty index consists of a large collection of macroeconomic and financial data. The volatility-based index is the VIX from the US stock exchange. Table 6 showcases the different indices and the corresponding labels:

Table 5

Top 10 words associated with uncertainty with corresponding rank, as evaluated by the attention layer of the proposed classifier.

2006		**2007**		**2008**		**2009**		**2010**	
subject	120	federal	134	presidential	200	presidents	88	presidential	110
federal	112	subject	84	economic	72	presidential	85	subject	90
government	69	newspaper	72	subject	69	subject	79	federal	78
economics	65	economics	61	crisis	54	stimulus	68	republican	69
newspaper	62	republican	48	republican	54	recession	68	newspaper	68
economic	45	budgets	45	federal	53	economic	62	recession	57
republican	41	presidential	44	budgets	45	crisis	60	economic	55
presidents	37	economic	40	newspaper	44	bailouts	56	presidents	54
english	32	government	38	presidents	42	newspaper	52	legislation	52
legislation	32	presidents	36	bailouts	42	federal	52	unemployment	50
2011		**2012**		**2013**		**2014**		**2015**	
presidential	118	presidential	196	presidential	93	american	73	rates	98
subject	84	subject	101	subject	84	america	48	american	71
newspaper	74	cliff	79	federal	63	republican	38	trump	48
debt	72	republican	74	newspaper	60	americans	38	rate	42
republican	70	presidents	52	ceiling	58	legislation	36	americans	42
recession	57	recession	50	republican	54	economic	33	america	40
federal	55	economics	46	recession	50	rates	32	federal	32
presidents	48	fiscal	46	cliff	49	federal	31	democrats	25
ceiling	46	economic	43	debt	48	newspaper	25	april	24
economics	46	federal	41	economics	43	subject	24	republican	22
2016		**2017**		**2018**		**2019**			
trump	343	trump	679	trump	774	trump	389		
brexit	153	rates	45	brexit	46	brexit	149		
americans	57	americans	44	tariffs	43	american	48		
american	49	american	37	american	38	america	44		
rates	49	america	26	april	22	tariffs	36		
america	41	legislation	23	california	19	rates	36		
rate	31	republican	21	rates	15	americans	23		
april	23	ms	14	americans	15	rate	19		
democrats	22	english	14	republican	14	republican	18		
republican	15	federal	12	immigration	13	true	15		

Table 6

Uncertainty proxies.

Name	Label	Source
BBD method on our data	BBD	own data
GPT-2 DNN	GPT-2	own construction
Total Uncertainty	S	Ozturk and Sheng (2018)
Real Uncertainty (h=1)	M	Jurado et al. (2015)
Stock market volatility	V	Bloom (2009)

4.3.1. Co-movement between uncertainty proxies

We start our economic analysis by looking at descriptive statistics of the uncertainty proxies in Table 7. Naturally, they need to be standardized for visual comparisons due to the varying value ranges. The kurtosis of all proxies except GPT-2 exceeds the value of the normal distribution, meaning that four out of five proxies show considerably high peaks. Stock market volatility and macro-based uncertainty include the highest peaks. Additionally, all proxies except GPT-2 are right-skewed, providing further evidence of relatively high values included in most proxies. GPT-2 is the closest to a normal distribution.

Table 7

Summary Statistics from Jan 2006 until Sept 2017.

Variable	Mean	Std. Dev	Min	Max	Skew	Kurt
BBD	94.95	47.51	17.56	248.77	1.29	4.57
GPT-2	92.56	35.40	29.35	175.38	0.14	1.98
S	0.43	0.31	0.11	1.23	1.30	3.59
M	0.64	0.05	0.58	0.83	2.08	7.13
V	19.44	9.18	10.26	62.64	2.33	9.66

We further look at correlations between the different proxies to identify potential groups. Figure 3 showcases the correlations among the different proxies, clustered by proximity:

We can see two distinct clusters formed by the newspaper-based proxies and all other proxies. The similarity among the other group of proxies seems to be higher than the similarity between the two economic policy uncertainty proxies. The EPU indices seem to be negatively correlated with the other group of uncertainty proxies. A negative correlation is counter-intuitive because it implies that if uncertainty measured by one group increases, the other group will decrease. To better understand this finding, we plot the different time series.

Figure 4 shows the five different proxies in one time series plot. We can see two major patterns: the group of indices that is not based on newspaper data has its peak around 2009 during the GFC and otherwise does not have any

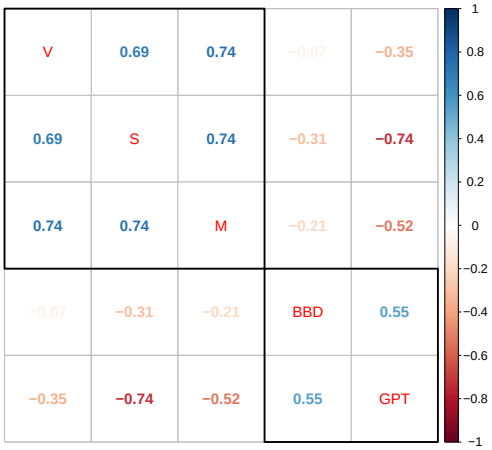

Figure 3: Pearson Correlation between the different uncertainty proxies from Jan 2006 until Sept 2017.

prominent peaks, while the newspaper-based indices have several.

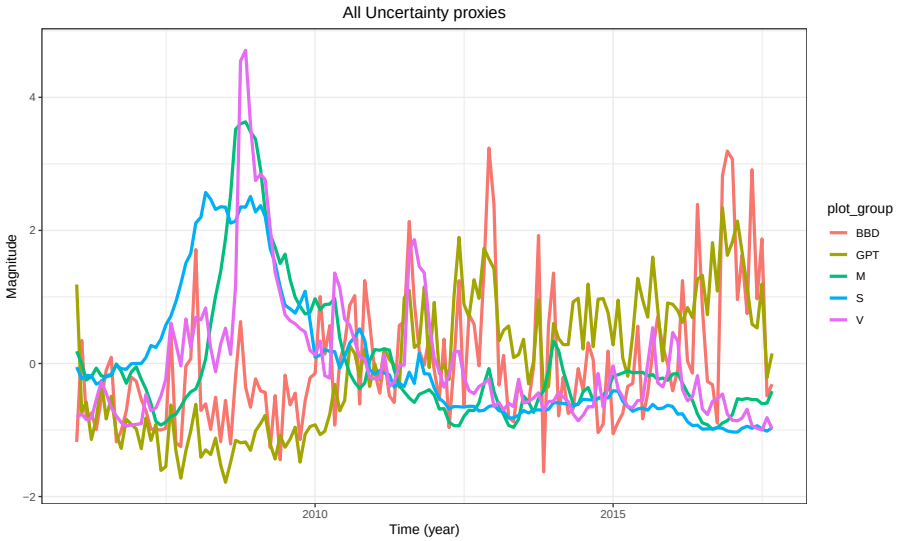

Figure 4: Time series plot of all uncertainty proxies from Jan 2006 until Sep 2017.

BBD and GPT-2 in Figure 4 have a relatively high variance. They also move up during all the times one would expect uncertainty to increase. Both indices are quite similar. The BBD-based index behaves slightly differently during the GFC from 2009 to 2012; otherwise, our GPT-2 index and the BBD index move up during all major events related to high uncertainty.

Survey uncertainty is especially visible during the GFC in 2009. Towards the end of the sample period, the uncertainty indicated by this proxy seems to fade out. For macro uncertainty, we obtain a similar picture: it reaches its

maximum during the GFC and shows comparatively little movement later on. Stock market volatility exhibits more variation than a survey- or macro-uncertainty but also peaks during 2009. It remains relatively smooth with a small peak during the European sovereign debt crisis in 2014.

To sum up, the uncertainty proxies based on macro-data, surveys, and volatility show similar behavior, potentially stemming from relying on all individual information in the economy that is available to individual agents before an uncertainty shock hits. BBD and GPT-2 show very different behavior from the other group of proxies. They have the highest variation among all uncertainty proxies and also the largest number of peaks. Instead of relying on an individual information, they are based on newspapers that already contain aggregated information.

The higher movement of the newspaper-based indices might indicate that these indices capture fast-moving uncertainty in the economy better than the other proxies, that mainly move during a small number of massive shocks.

4.3.2. Interaction with real economic variables

We investigate if GPT-2 is better than BBD at predicting the movements of the economy and capturing the change in newspaper vocabulary. Therefore, we measure how BBD and GPT-2 correlate with different real economic variables. In general, the latter should be negatively correlated with the uncertainty proxies, so that when uncertainty increases, the affected variables decrease. Based on theoretical literature in economics (Arellano et al., 2019; Bernanke, 1983), uncertainty affects the variables in Table 8. To account for possible non-stationarity of the variables we try the following specifications: level, difference, hp-filter, residuals after fitting an ARIMA model. Since the level-level specification always shows the highest correlation, we decided to display only those results.

Table 8
Macroeconomic and Financial variables.

Name	Label
S&P 500	SP
Employment (manufacturing)	EM
Industrial production	IP

We explore the relationship between the time series with Pearson correlation. Since we do not know which lag yields the most substantial relationship between uncertainty and economic variables, we explore this dynamic by showing the correlations across the range from t_0 up to t_{0+k}, where $k = 36$.

Figures 5, 6 and 7 show the correlation between uncertainty proxies and industrial production, employment and the stock market respectively.

The correlation with industrial productivity is positive, steadily increasing and attains it's maximum at a log of 25 months. The correlation with industrial employment is positive and has it's maximum at a lag of 0. From there on, it is slowly decreasing, reaching a correlation of approximately 0 around lag 36. The correlation with the stock market

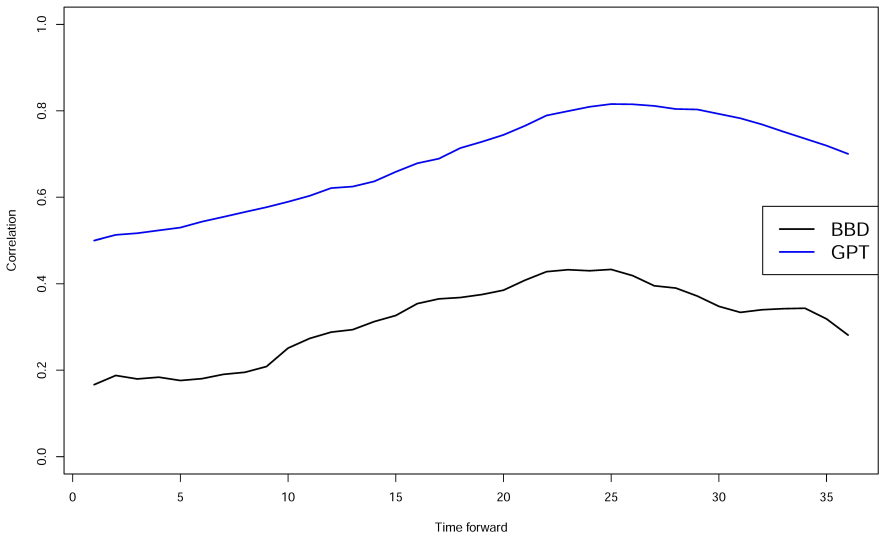

Figure 5: Correlation of uncertainty proxies with industrial production.

is positive. While GPT-2 has a relatively stable relation to the stock market over time, the correlation between BBD and the stock market fluctuates more.

Generally, it seems that the connection between GPT-2 and economic activity is stronger than the connection between BBD and economic activity since the first one always results in higher correlation. We find positive correlations, which are not in line with theory but this could be expected since our category of uncertainty proxies showed very different properties from all other categories of uncertainty proxies.

4.3.3. Forecasting performance

In this Section, we evaluate the potential of our newly created index for forecasting, following the practice of Claveria et al. (2007); D'Amuri and Marcucci (2017); Tarassow (2019). We are forecasting five different variables with an ARIMAX model, where either our GPT-2 index or the BBD index is added as an exogenous variable for forecasting. The forecasting is done for rolling windows of 18, 12, and 6 months during the three different periods explained in the previous subsection and performed with the *auto.arima()* function from the forecast package in R by Hyndman and Khandakar (2007).

For each period and variable, we obtain one forecast where the model is augmented by the BBD labels and one forecast where the model is augmented by the predictions of the GPT-2 DNN model. Tables 9 and 10 exhibit the RMSE for all forecasts. Additionally, we perform Diebold-Mariano tests to identify superior forecasting performance among the ARIMAX models.

For the whole sample period, the model, including the BBD index, generally seems to result in lower RMSE for six months and the government bond yield. This difference in RMSE is only statistically significant for the forecasts

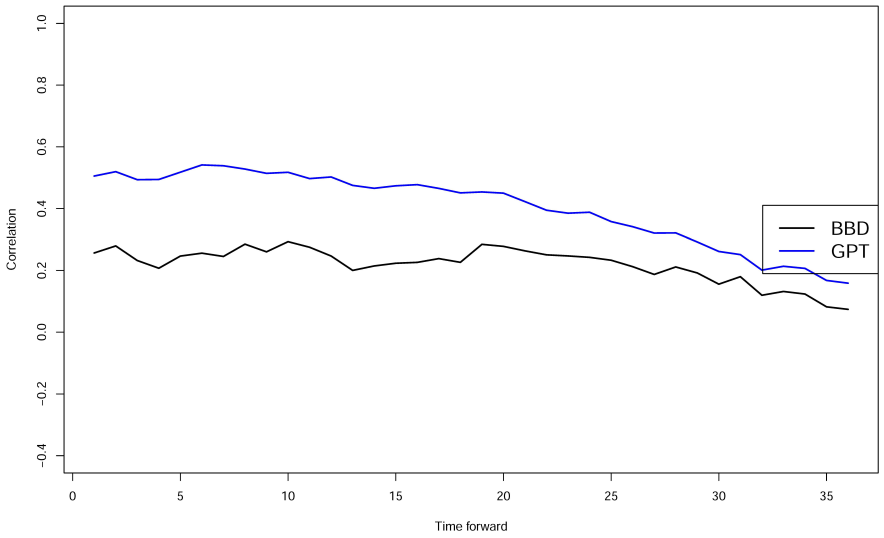

Figure 6: Correlation of uncertainty proxies with industrial employment.

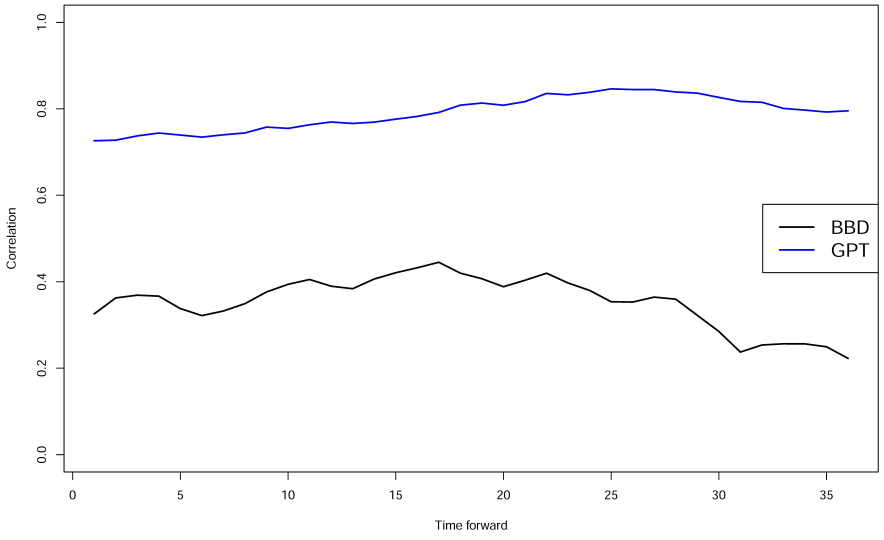

Figure 7: Correlation of uncertainty proxies with the stock market.

Table 9

ARIMAX forecasts for Period 1.

Model	S&P 500	Fed funds	Empl.	Ind. prod.	Bonds
Window of 18 months					
BBD	0.048	0.026	0.002	0.009	0.171*
GPT-2	0.047*	0.026	0.002	0.009	0.178
Window of 12 months					
BBD	0.050	0.029	0.002	0.008	0.182
GPT-2	0.050	0.030	0.002	0.009	0.201
Window of 6 months					
BBD	0 .051**	0.039	0.006	0.010	0.242
GPT-2	0.068	0.036	0.007	0.012	0.250

This table shows RMSE for rolling window ARIMAX models with BBD and GPT-2 as external regressors. Stars indicate significance levels of Diebold-Mariano Tests for higher forecast accuracy: ** = 0.05; * = 0.1.

of the stock market. GPT-2 provides lower RMSE only when forecasting the federal funds rate with a rolling window of six months. Otherwise, the forecasts are very similar.

Since our goal is to investigate whether our index can deal better with the change in the vocabulary used by the newspapers, we carry out the same forecasting exercise for the previously defined Periods 2 and 3. The results can be found in Table 10.

Periods 1 and 2 exhibit a similar pattern. BBD provides forecasts of higher accuracy for the stock market at a rolling window of 6 months, as well as the forecasts of government bond yields at a window of 18 months. GPT-2 only yields lower RMSE for the federal funds rate at a window of six months. For all other variables and window sizes, there is no statistical difference in RMSE between the two models.

For the Period 3, RMSE is generally much higher. BBD does not yield lower RMSE anymore. Instead, GPT-2 shows lower RMSE for all variables for a window of six months. The difference in RMSE is statistically significant for forecasts of the federal funds rate and industrial production. For the longer windows, both models show similar forecast accuracy.

To sum up, for P1 and P2, BBD generally yields forecasts with lower RMSE, even though there is rarely a statistically significant difference between the two models. In Period 3, when a change of newspaper agenda occurred, our model provides lower RMSE and more accurate forecasts for two out of five variables. This serves as evidence that a DL-NLP-based index can better deal with changing newspaper agendas over time.

Table 10

ARIMAX forecasts for Period 2 and 3.

Model	S&P 500	Fed funds	Empl.	Ind. prod.	Bonds	S&P 500	Fed funds	Empl.	Ind. prod.	Bonds
Window of 18 months P2						Window of 18 months P3				
BBD	0.052	0.028	0.002	0.001	0.200*	0.856	0.240	0.009	0.141	1.559*
GPT-2	0.052	0.028	0.002	0.001	0.204	0.856	0.239	0.009	0.141	1.577
Window of 12 months P2						Window of 12 months P3				
BBD	0.056	0.032*	0.002	0.008	0.212	0.796	0.228	0.096	0.129	1.487
GPT-2	0.055	0.034	0.002	0.010	0.235	0.795	0.230	0.096	0.129	1.438*
Window of 6 months P2						Window of 6 months P3				
BBD	0.055**	0.043	0.007	0.012	0.277	0.745	0.221	0.102	0.116	1.443
GPT-2	0.079	0.038	0.009	0.014	0.285	0.735	0.211**	0.101	0.113*	1.433

This table shows RMSE for rolling window ARIMAX models with BBD and GPT-2 as external regressors. Stars indicate significance levels of Diebold-Mariano Tests for higher forecast accuracy: ** = 0.05; * = 0.1.

5. Conclusion

We offer a DL-NLP-based method for the quantification of economic policy uncertainty. The method is applied to the corpus of articles from ten major USA newspapers from 01 Jan 2006 to 30 Apr 2019. The predictive performance of our model surpassed the benchmarks like Support Vector Machine or Random Forest with an AUC of 0.96 and an F1-score of 0.65. The model remained robust in 10-fold cross-validation.

Our method offers high interpretability and adaptability, which was demonstrated by the analysis of the top ten words responsible for EPU over time. We exposed a definite change of agenda in the newspaper articles. The first part of the sample, from Jan 2006 until Dec 2014, did not feature the word ”trump”. Starting in Jan 2015 until the end of our sample in Apr 2019, the word ”trump” always featured in the top ten. These shifts show the necessity to adapt to changing political and economic trends when trying to capture economic uncertainty from newspaper articles.

By investigating the correlations between our uncertainty proxies and economic activity, we provided evidence that machine learning succeeds in extracting more relevant information from newspaper articles than manually determined keywords. This is illustrated by the higher correlation between the DL-NLP-based index and economic activity across all selected variables.

With our forecasting experiment, we showed that during the later period, forecasting accuracy reduced drastically. Our uncertainty index based on DL-NLP had superior forecasting ability for two out of five variables and resulted in lower RMSE for all variables. In the earlier period, none of the two models provided higher accuracy for four out of five variables. This way, the proposed method proved its suitability to deal with the change in newspaper agenda better than the methodology of Baker et al. (2016).

Our approach shows pathways towards capturing economic policy uncertainty over long periods while keeping

track of changes in the way that news and uncertainty are reported. Two recent examples that changed newspaper reporting are the Trump presidency and the recent COVID-19 pandemic. The approach might prove especially useful for governments and institutions in countries with scarce, timely information sources on the level of uncertainty in the economy as newspaper articles are widely available over time and therefore represent a feasible alternative data source to assess economic policy uncertainty.

References

Alexopoulos, M., Cohen, J., 2015. The power of print: Uncertainty shocks, markets, and the economy. International Review of Economics and Finance 40, 8–28.

Arellano, C., Bai, Y., Kehoe, P.J., 2019. Financial frictions and fluctuations in volatility. Journal of Political Economy 127, 2049–2103.

Athiwaratkun, B., Stokes, J.W., 2017. Malware classification with lstm and gru language models and a character-level cnn, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2482–2486.

Azqueta-Gavaldón, A., 2017. Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. Economics Letters 158, 47–50.

Bachmann, R., Elstner, S., Sims, E.R., 2013. Uncertainty and Economic Activity : Evidence from Business Survey Data. American Economic Journal: Macroeconomics 5, 217–249.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., 1989. A tree-based statistical language model for natural language speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 37, 1001–1008.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131, 1593–1636.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155.

Bernanke, B.S., 1983. Irreversibility, uncertainty, and cyclical investment. The Quarterly Journal of Economics 98, 85–106.

Bloom, N., 2009. The Impact of Uncertainty Shocks. Econometrica 77, 623–685.

Bobasu, A., Geis, A., Quaglietti, L., Ricci, M., 2020. Tracking global economic uncertainty: implications for global investment and trade. Economic Bulletin Boxes .

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 .

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 .

Chorowski, J., Bahdanau, D., Cho, K., Bengio, Y., 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. arXiv preprint arXiv:1412.1602 .

Christiano, L.J., Motto, R., Rostagno, M., 2014. Risk shocks. American Economic Review 104, 27–65.

Claveria, O., 2019. Forecasting the unemployment rate using the degree of agreement in consumer unemployment expectations. Journal for Labour Market Research 53, 1–10.

Claveria, O., Pons, E., Ramos, R., 2007. Business and consumer expectations and macroeconomic forecasts. International Journal of Forecasting 23, 47–69.

Dai, A.M., Le, Q.V., 2015. Semi-supervised sequence learning, in: Advances in Neural Information Processing Systems, pp. 3079–3087.

D'Amuri, F., Marcucci, J., 2017. The predictive power of google searches in forecasting us unemployment. International Journal of Forecasting 33, 801–816.

Deng, L., Yu, D., 2014. Deep learning: methods and applications. Foundations and Trends® in Signal Processing 7, 197–387.

Ghirelli, C., Perez, J.J., Urtasun, A., 2019. A new economic policy uncertainty index for spain. Economics Letters , 64–67.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE. pp. 80–89.

Glas, A., 2019. Five dimensions of the uncertainty-disagreement linkage. International Journal of Forecasting 36, 607–627.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning: adaptive computation and machine learning. MIT press.

Graves, A., Mohamed, A.r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, in: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE. pp. 6645–6649.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 .

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349, 261–266.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 .

Hyndman, R.J., Khandakar, Y., 2007. Automatic time series for forecasting: the forecast package for R. 6/07, Monash University, Department of Econometrics and Business Statistics ....

Jurado, K., Ludvigson, C.S., Ng, S., 2015. Measuring Uncertainty. American Economic Review 105, 1177 – 1216.

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences, in: 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, Association for Computational Linguistics (ACL). pp. 655–665.

Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M.C., Johnson, J.E., 2020. Can deep learning predict risky retail investors? a case study in financial risk behavior forecasting. European Journal of Operational Research 283, 217–234.

Kimball, M.S., 1990. Precautionary Savings in the Small and in the Large. Econometrica 58, 53–73.

Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research 259, 689–702.

Larsen, V., 2017. Components of uncertainty. Norges Bank Working Paper Series .

Leduc, S., Liu, Z., 2016. Uncertainty shocks are aggregate demand shocks. Journal of Monetary Economics 82, 20–35.

Liu, F., Zheng, J., Zheng, L., Chen, C., 2020. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. Neurocomputing 371, 39 – 50.

Liu, Y., Sheng, X.S., 2019. The measurement and transmission of macroeconomic uncertainty: Evidence from the us and bric countries. International Journal of Forecasting 35, 967–979.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Ozturk, E.O., Sheng, X.S., 2018. Measuring global and country-specific uncertainty. Journal of International Money and Finance 88, 276–295.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 .

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1, 9.

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 2673–2681.

Sill, K., 2014. Forecast disagreement in the survey of professional forecasters. Business Review Q 2, 15–24.

Soric, P., Lolic, I., 2017. Economic uncertainty and its impact on the croation economy. Public Sector Economics 4, 443–477.

Tarassow, A., 2019. Forecasting us money growth using economic uncertainty measures and regularisation techniques. International Journal of

Forecasting 35, 443–457.

Thorsrud, L.A., 2018. Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. Journal of Business and Economic Statistics , 1–17.

Tobback, E., Naudts, H., Daelemans, W., Junqué de Fortuny, E., Martens, D., 2018. Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting 34, 355–365.

Tripier, F., 2019. Assessing the cost of uncertainty created by brexit. EconPol opinion .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems, pp. 5998–6008.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480–1489.

Yin, W., Kann, K., Yu, M., Schütze, H., 2017. Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923 .

Zalla, R., 2017. Economic policy uncertainty in ireland. Atlantic Economic Journal 2, 269–271.

Zhang, Z., Robinson, D., Tepper, J., 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European semantic web conference, Springer. pp. 745–760.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B., 2016. Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pp. 207–212.