DISCUSSION PAPER SERIES

DP17410

Measuring Brexit Uncertainty: A Machine Learning and Textual Analysis Approach

Wanyu Chung, Duiyi Dai and Robert Elliott

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

ISSN 0265-8003

Measuring Brexit Uncertainty: A Machine Learning and Textual Analysis Approach

Wanyu Chung, Duiyi Dai and Robert Elliott

Discussion Paper DP17410
Published 27 June 2022
Submitted 24 June 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre’s research programmes:

- International Trade and Regional Economics

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Measuring Brexit Uncertainty: A Machine Learning and Textual Analysis Approach

Abstract

In this paper we develop a series of Brexit uncertainty indices (BUI) based on UK newspaper coverage. Using unsupervised machine learning (ML) methods to automatically select topics, our main contribution is to generate timely and cost-effective indicators of uncertainty. In further analysis we are able to distinguish Brexit related uncertainty from the uncertainly due to COVID-19. Our indices can be used to investigate Brexit-related uncertainties across different policy areas.

JEL Classification: D80, F50, E66

Keywords: Brexit, uncertainty, Machine Learning

Wanyu Chung - w.chung@bham.ac.uk
University of Birmingham and CEPR

Duiyi Dai - dxd048@student.bham.ac.uk
University of Birmingham

Robert Elliott - r.j.elliott@bham.ac.uk
University of Birmingham

Measuring Brexit Uncertainty: A Machine Learning and Textual Analysis Approach[*]

Wanyu Chung[†]
University of Birmingham and CEPR

Duiyi Dai
University of Birmingham

Robert J R Elliott
University of Birmingham

June 24, 2022

Abstract

In this paper we develop a series of Brexit uncertainty indices (BUI) based on UK newspaper coverage. Using unsupervised machine learning (ML) methods to automatically select topics, our main contribution is to generate timely and cost-effective indicators of uncertainty. In further analysis we are able to distinguish Brexit related uncertainty from the uncertainly due to COVID-19. Our indices can be used to investigate Brexit-related uncertainties across different policy areas.

JEL Codes: D80, F50, E66
Keywords: Brexit; Uncertainty; Machine learning

[*] We are grateful for financial support from Research England for the QR Enhancing Research and Knowledge Exchange Policy Support Fund (2022). Any views expressed are solely those of the authors. We would like to thank Diane Coyle, Christoph Gortz, Michael McMahon, Paul Mizen, Oleksandr (Sasha) Talavera, and seminar participants at the Public Policy Exchange's policy forum "UK–EU Trade: Improving Administrative Procedures & Supporting UK Businesses Trading with Europe Post-Brexit," Peter Sinclair Memorial Conference (University of Birmingham and University of Oxford), University of Aberdeen, Annual Southern PhD Economics Conference, and ESRC MGS DTP Annual Conference.

[†] Chung: Assistant Professor in Economics, Department of Economics, University of Birmingham, Birmingham B15 2TY, UK; and CEPR. Email: w.chung@bham.ac.uk; Dai: PhD candidate, Department of Economics, University of Birmingham. Email: dxd048@student.bham.ac.uk (corresponding author); Elliott: Professor of Economics, Department of Economics, University of Birmingham. Email: r.j.elliott@bham.ac.uk.

1 Introduction

The term "Brexit" was first used in May 2012, eight months before the then British Prime Minister David Cameron announced he would hold a referendum on whether the UK should leave the EU.[1] The term "Brexit" is now in popular usage especially after the largely unanticipated result that the UK was to leave the EU. An immediate reaction to the vote was a substantial and protracted increase in uncertainty that impacted different aspects of the UK economy with some sectors and groups of people more heavily impacted than others. For example, the Bank of England (2019) points out that Brexit uncertainty has driven delays in firm investment and depressed productive capacity.[2]

The purpose of the paper is to construct aggregate and topic-specific news-based indices of Brexit uncertainty for the UK economy using techniques from computational linguistics and can be updated in close to real time. Our Brexit uncertainty indices (BUIs) are derived from the frequency of relevant news coverage from eleven leading UK newspapers. A growing body of literature has documented the negative effect of uncertainty on macroeconomic activity.[3] Explanations for such negative effects include that firms will often employ a wait-and-see strategy when there is reduced visibility of the future (Bernanke, 1983), increased financing costs (Gilchrist et al., 2014), and search frictions in the labor market as well as nominal rigidities (Leduc and Liu, 2016). The definition of uncertainty used in macroeconomics is the conditional volatility of economic or policy shocks that cannot be predicted (Jurado et al., 2015). In the context of this paper, the term Brexit uncertainty captures a range of possible uncertainties faced by the UK economy that were triggered by the decision to withdraw from the EU and are subsequently reflected in UK newspaper coverage.

The contribution of the paper is threefold. First, we provide a real-time and cost effective method for measuring historic and ongoing Brexit uncertainty. Recent studies have tended to rely on aggregate Brexit uncertainty proxies that encompass parts of the Brexit process, for example, a proxy for pre-referendum uncertainty using public expectations of the Brexit referendum outcome in prediction markets (Graziano et al., 2020, 2021), and for post-referendum uncertainty using a firm-level survey (Bloom et al., 2019). Our news-based index that starts in 2013, provides consistent and comparable indices before, during, and after the referendum. Second, in addition to an aggregate BUI, for the first time to the best of our knowledge, we provide indices that capture the uncertainty associated with a number of specific topics and sectors. Topics include trade policy, immigration, Northern Ireland, supply chains, energy & climate, employment and the broader macro-economy. Third, we disentangle the uncertainty induced by COVID-19 from Brexit for both the aggregate BUI

[1] "The rise of the word Brexit," BBC news 2016.

[2] See "Monetary Policy Report - November 2019: In focus - Uncertainty and Brexit" published by the Bank of England.

[3] See, for example, Bloom (2009), Fernández-Villaverde et al. (2011), Hassan et al. (2019), Bloom et al. (2018), Arellano et al. (2019), Bachmann et al. (2013), Basu and Bundick (2017), and Fajgelbaum et al. (2017).

and the topic based indices. This is important as the UK economy experienced and is experiencing considerable uncertainty as a result of the pandemic (Altig et al., 2020). This approach allows us to quantify how the COVID and Brexit uncertainly interact and hence demonstrate how COVID may have masked or exaggerated alternative measures of Brexit related uncertainty.

Our methodological approach is to use machine learning algorithms to identify the news coverage of different aspect of Brexit uncertainty. For example, our aggregate BUI captures the frequency of newspaper articles that contain the word "Brexit", a word suggesting "uncertainty", and a word indicating the country "UK". To do this we adapt the Word2Vec algorithm developed by Mikolov et al. (2013) to find semantically similar words to "uncertainty" and "UK" in a context where Brexit uncertainty is being described. The algorithm is able to capture semantic similarities between words by learning from a sub-sample of the news data. To this end, inputting "uncertain" and "UK" as seed words, the algorithm outputs semantically related words which we then prune using informed judgment to expand the word sets.

To capture how Brexit related uncertainty impacts different topics, we rely on the Latent Dirichlet allocation (LDA) model, a probabilistic topic model, introduced by Blei et al. (2003) to decompose the content of articles to obtain the probability or the "share" of each article that is related to each Brexit topic, given that multiple topics could be covered in any one article. For this analysis, all articles containing the word "Brexit" and at least one word in the "uncertain" and "UK" sets are curated, processed, and composed into a corpus. This model automatically learns the corpus and extracts multiple topics in the form of probability distributions over words. For each article, a probability distribution over topics is obtained after fitting the LDA model. We therefore obtain the proportion of newspaper coverage of our corpus linked to each topic at each point in time. The subsequent time-varying proportions then constitute topic-specific BUIs. As a result, our topic-specific BUIs capture a series of uncertainty issues associated with the Brexit process. These issues include supply chain concerns, the UK-France dispute over fishing rights before the end of transition period and Northern Ireland close to the time when the Northern Ireland Protocol was enacted.

Although measuring Brexit related uncertainly using our machine learning approach was relatively straight forward for the pre- and early post referendum result periods, the onset of the COVID-19 pandemic presented a major challenge given the uncertainty that the global pandemic induced. Our solution is to disentangle the uncertainty driven by Brexit from that driven by COVID for the post-2020 period. Post 2020, both COVID and Brexit contribute to the uncertainty faced by the UK with many of the same issues being impacted such as international trade, labour shortages, and supply chains.[4] An important concern

[4] A number of newspaper articles discuss the relationship between COVID and Brexit uncertainty. For example, "COVID pandemic masks Brexit impact on UK economy" published by Financial Times in 2021 and "Impact of Brexit on economy 'worse than COVID'" published by BBC in 2021. On average, over three-quarters of businesses viewed COVID as their top source of uncertainty after March 2020. Data from

is therefore the extent to which the COVID-19 pandemic has a distortionary effect on our BUI indices. To this end, we re-compute our BUIs excluding articles that mention COVID. This approach provides a lower bound for our BUIs. In practical terms, our approach is cost-effective in both human and financial terms (there is no costly survey to manage), and it enables us to examine both the pre- and post-referendum periods, and is not subject to the survey sample size and the response rate of those asked to complete a survey. Moreover, our approach means we can easily examine different Brexit related topics and hence provide a level of disggregation not previously analysed.

The next stage of our analysis is to provide a detailed validation process to ensure we are accurately capturing Brexit related uncertainty. First, we investigate whether our aggregate BUI demonstrates strong co-movements with the well known survey-based Brexit uncertainty series from Bloom et al. (2019) which captures the proportion of managers who rate Brexit as the top source of uncertainty for their companies. We also compare our index against a measure of economic policy uncertainty from Baker et al. (2016). In other analysis we investigate the role of political leanings in the newspapers selected and finally compare trends in our BUIs and UK stock market volatility.

Our paper relates to two main strands of literature. The first strand of literature looks explicitly on the effects of Brexit uncertainty. This uncertainty has been shown to impact investment and productivity (Bloom et al., 2019), employment (Javorcik et al., 2020), the EU-UK bilateral trade in goods (Crowley et al., 2018; Douch et al., 2020; Graziano et al., 2021) and trade in services (Ahmad et al., 2020). Some of these studies also look at how uncertainly impacts certain sectors. For example, Crowley et al. (2018), Javorcik et al. (2020), and Douch et al. (2020) examine the cross-sectional variation in trade policy-related Brexit uncertainty utilizing the contingent gaps in trade terms for different Brexit outcomes. Such measures, however, do not capture the evolution over time although Bloom et al. (2019) do show how uncertainly evolves over time by asking managers the same question each year and recording the number of managers who report Brexit as a major source of uncertainty for their businesses using the Decision Maker Panel (DMP).

The second strand of the literature conducts text analysis in economics settings in which news-based measures of uncertainty have received great attention. The most widely used text-based uncertainty metric is the so-called dictionary method, which identifies a set of terms of interest for researchers and then calculates the frequency of those terms across text corpus to generate an uncertainty index. Examples include Baker et al. (2016) for economic policy uncertainty, Caldara and Iacoviello (2022) for geopolitical risk, Husted et al. (2020) for monetary policy uncertainty, and Caldara et al. (2020) for trade policy uncertainty. Leveraging machine learning tools, our method seeks to reduce subjective identification of terms and to allow for a larger feature space, i.e., accounting for many more words rather than merely those in the dictionary. Further, in a news-based dictionary method, the small-

Decision Maker Panel (DMP): https://decisionmakerpanel.co.uk/.

est unit is typically an article. More specifically, each qualifying article contributes equally to the topic-specific uncertainty measurement, despite the fact that the fraction of content concerning the topic of interest varies considerably from article to article. In this paper, we employ the LDA model to decompose articles into a distribution of topics, and weight each article accordingly. Hansen et al. (2018), to our knowledge, is the first to introduce the LDA model into economics research.[5] Previous studies also use the LDA model to examine different types of uncertainty, for example, economic policy uncertainty (Azqueta-Gavaldón, 2017), and general business uncertainty (Larsen, 2021). In addition to being the first to introduce the LDA model into a Brexit uncertainty context we also rely on the Word2Vec model to expand searching term sets. By doing so, we reduce the subjectivity in the identification of term sets. Previous applications of this model include Burn et al. (2019), and Davis et al. (2020). There is also a growing literature on forecasting with text (see Kalamara et al. (2022) for a discussion).

To briefly summarise our results we find that the main BUI spikes occurred around the Brexit referendum, the three failed meaningful votes, when the Brexit withdrawal agreements were rejected in the House of Commons in early 2019, and the period just before the final agreement was reached. In terms of average values, Brexit uncertainty was high between the announcement of the referendum date (February 2016) and the UK's formal departure from the EU (December 2019). Our results also show that even after the Brexit deal was reached, Brexit uncertainty did not fall significantly (post-2021) with a magnitude roughly four-fifths of that during the transition period. By taking into account of COVID uncertainty we show that Brexit uncertainty was exaggerated by as much as 1.5 times in the post-2020 period due to the pandemic, with the most pronounced magnification effect related to employment (Covid effect accounting for 77.9%), government spending & budgets (74.3%), and supply chains (69.0%). Results suggest that fishing rights and Northern Ireland were the least affected by a COVID effect accounting for 29.1% and 30.1%, respectively.

Validation exercises reveal a correlation between our aggregate BUI and the DMP index of 0.82 rising to 0.93 when we drop the first two years of their index due to lower number of panel members and survey respondents (see Bloom et al. (2019)). Our lower bound BUI (excluding COVID pandemic effects) results in an even higher overall correlation of 0.86. Our aggregate BUI also has a positive correlation with the UK economic policy uncertainty (EPU) index constructed by Baker et al. (2016) which recorded a correlation of 0.73 for the post-2016 period, during which Brexit emerges as the main source of economic policy uncertainty in the UK economy. Finally, in a further robustness check we find significant co-moving trends between currency-related BUI and UK stock market volatility after the referendum.

The remainder of this paper will proceed as follows. Section 2 presents the data and de-

[5] Other applications of this model in economics settings include, for example, Hansen and McMahon (2016), Hansen et al. (2019), Larsen and Thorsrud (2019), Bybee et al. (2020), and Larsen et al. (2021).

scribes the methodological approach. Section 3 presents our different indices and discusses the policy implications. Section 4 describes and shows the results of our validation process. Section 5 concludes.

2 Measuring Brexit Uncertainty

In this section we describe in detail the data and methodology used to measure aggregate and topic specific Brexit uncertainty. Put simply, our aggregate and topic specific Brexit uncertainty indices are constructed based on the frequency of newspaper coverage devoted to Brexit uncertainty. In Section 2.1 we describe the news data and the cleaning process.

In terms of methodology, two machine learning (ML) models are used to assist in identifying news content. A flowchart that illustrates our approach is shown in Figure 1. First, we filter out articles that touch on the topic of Brexit uncertainty from the mass of news, based on the inclusion or not of the word "Brexit" and words referring to “uncertainty” and “UK.” The Word2Vec model helps by outputting the groups of words that indicate “uncertainty” and “UK.” A detailed description of the Word2Vec model can be found in Section 2.2 where we explain the construction of the aggregate BUI.

The second ML algorithm, the LDA model, serves to match the newspaper coverage to Brexit topics by identifying topics in the form of distributions over words, and subsequently, topic distribution over each news article in our news corpus. We explain the detailed model and how topic-specific BUIs are calculated in Section 2.3.

2.1 News Data Source and Pre-processing

Our BUIs are based on articles from eleven leading British newspapers which are *The Financial Times, The Times, The Sunday Times, The Daily Telegraph, The Daily Mail and Mail on Sunday, The Daily Express, The Guardian, The Mirror, The Sun, The Northern Echo*, and *The Evening Standard*. We deliberately choose the same newspaper list used by Baker et al. (2016) to quantify UK economic policy uncertainty (this also allows us to compare our results later in the paper). The articles of interest are collected from *Nexis* which is an online database that archives a broad range of news sources. We chose January 2013, when the Brexit referendum was first announced by the Prime Minister, as the starting point for our data collection. Prior to this, the term “Brexit” was not in popular usage. Data was collected up until April 2022.

Once the data has been collected it needs to be pre-processed and cleaned before it can be used with our ML algorithms. First, a pattern matching technique is employed to automatically extract publication date and the content of each news article from the downloaded

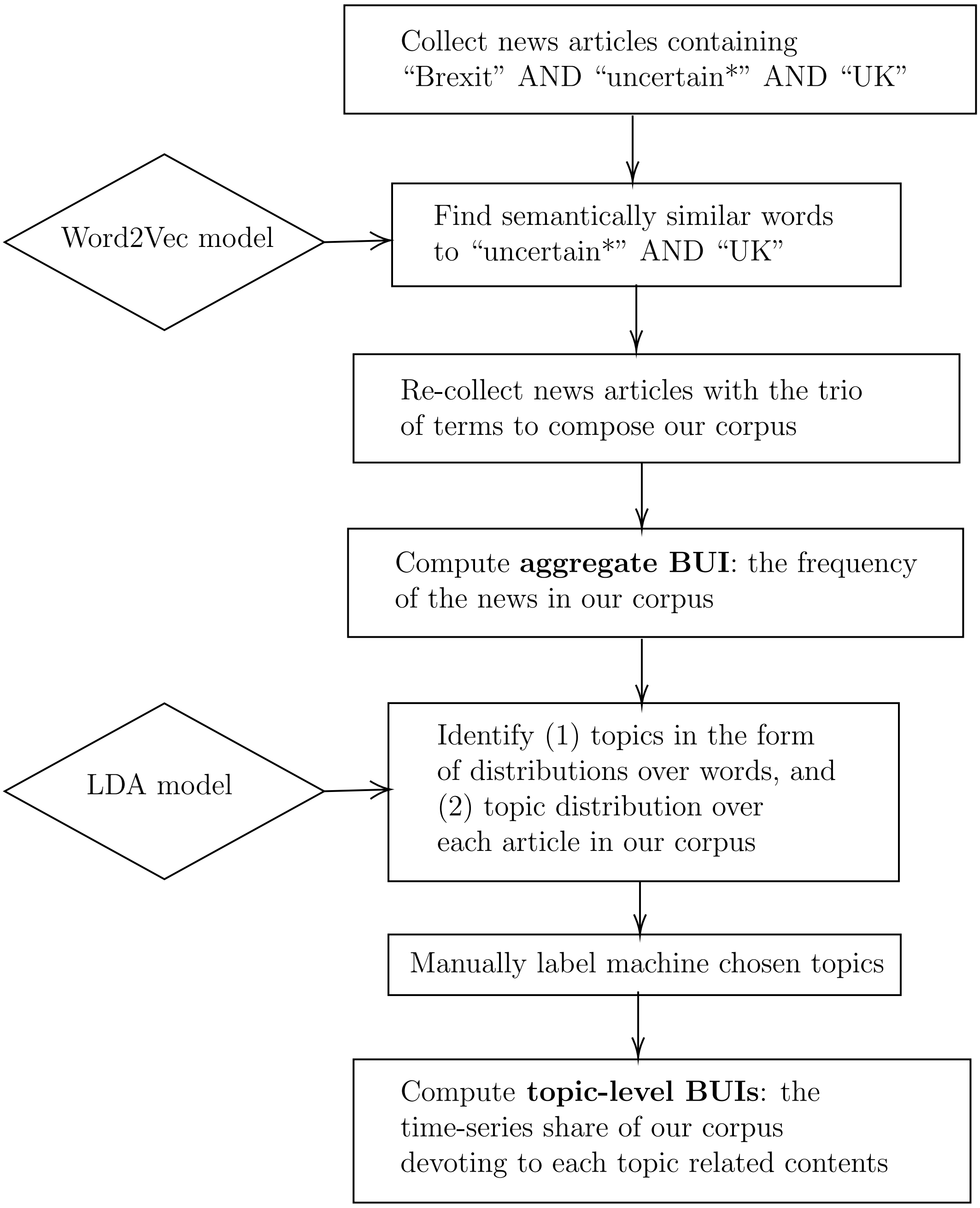

Collect news articles containing "Brexit" AND "uncertain*" AND "UK"
Word2Vec model
Find semantically similar words to "uncertain*" AND "UK"
Re-collect news articles with the trio of terms to compose our corpus
Compute **aggregate BUI**: the frequency of the news in our corpus
LDA model
Identify (1) topics in the form of distributions over words, and (2) topic distribution over each article in our corpus
Manually label machine chosen topics
Compute **topic-level BUIs**: the time-series share of our corpus devoting to each topic related contents

Figure 1: A flowchart to illustrate how we construct our measures of aggregate and topic specific BUIs using the two ML models, Word2Vec and LDA.

documents. In this case, we look at two common patterns. The first pattern is to locate and extract the news body by the fixed words or sentences that appear before and after the news body, that is, "Body" and "End of Document". The second pattern is to locate the common expression of publication date which is invariably in the order of the month, e.g., January, the one or two-digit date, e.g., 18, a comma or space, and then the four-digit year, e.g., 2013. We take the first eligible date in each document as the publication date.

The next step is to remove duplications by excluding those articles with identical body parts. Duplication is mostly attributed to various editions of the same piece of news. We then clean the news content data by converting uppercase letters to lowercase, removing stop words that do not carry information, e.g., “the,” “a” and “is,” punctuations, white spaces and numbers. The fourth step is tokenization, i.e., splitting the text data into individual words. Finally, the words are stripped down to their stems, for example, the words “manufacturing,” “manufacture,” “manufacturer” and “manufacturers” are all reduced to “manufactur”. The word stems are not necessarily English words.

2.2 Measuring Aggregate BUI: Expanding Searching Words with Word2Vec Model

The main assumption underpinning our analysis is that articles describing Brexit uncertainty in the UK contain the term “Brexit,” one or more terms indicating “uncertainty” and at least one term suggesting the country “UK.”[6] In the dictionary method, those semantically similar terms are defined manually. This process is subjective and, as such, it is difficult to obtain uniform term sets in multiple pieces of research which could potentially generate inconsistent results. Here, we draw on a ML algorithm that means subjective judgements in the identification of words in “uncertainty” and “UK” term sets is reduced.

More specifically, we adopt the Word2Vec model developed by Mikolov et al. (2013) to look for words that are semantically similar to “uncertain” and “UK.” We choose this technique for its relative simplicity and long-term good performance in ML applications. Around 10,000 news articles including “Brexit”, “uncertain*” and “UK” are collected and cleaned to train the model.[7]

The key to the model is word embedding, that is, word vectorization. The outputs from the model is word vectors that indicate the association between words by performing dimensionality reduction of vector space, in other words, mapping all words to low-dimensional vectors. In general, the simplest vectorization is to create a vector space in which each dimension corresponds to a unique word. Suppose that the corpus involves V unique words in total, then the vector space would have V dimensions. Word vectors would only have one non-zero valued element corresponding to its particular dimension, and the other elements would be 0, i.e., one-hot encoded vector. These vectors are orthogonal with zero cosine similarity between them, and thus, not conveying the similarity between words. Using the Word2Vec model, the vector space is reduced (to less than V), and similar words are assigned to vectors that are close on the vector space. The similarity between words can be

[6] Adding “EU referendum,” “European Union referendum,” “UK referendum,” “UK’s withdrawal,” “EU Exit,” “ EU withdrawal,” “leave the EU,” “exit from the EU,” “Exit the EU,” “Withdrawal from the EU,” to indicate Brexit does not change the resulting index. Indeed, the correlation is 1 when these terms are included.

[7] “*” means all the words with the root. For example, “uncertain*” includes “uncertain,” “uncertainty,” “uncertainties,” and “uncertainly.”

predicted based on the context in which the words appear, that is, the model defines words that appear in similar contexts as similar words.

We apply the model using a continuous bag-of-words (CBOW) architecture. In essence, the model with CBOW architecture aims to predict the center words, conditional on their context words. The algorithm works as a deep learning classification model, and self-generates the observation-label pairs, that is, the context-center word pairs. Given that the context of center word can be lengthy, it is necessary to set a window size C when training this model, signifying that the preceding and following C words of the center word are taken as the context. Suppose that the corpus involves V unique words in total, the input layer, then, consists of V-dimensional one-hot encoded vectors $(X_{-c}, \ldots, X_{-1}, X_1, \ldots, X_c)$ of contextual words with size $2C$. The output layer is the one-hot encoded vector of the center word X_0. In the process of predicting the central word or in the hidden layer of the neural network, the model computes a N-dimensional vector representation of it, in which the size of vector space N is pre-defined. Then, the distance, i.e., the cosine similarity, between all word vectors can be calculated.[8] These values range between -1 and 1, with larger values corresponding to greater similarity. The distance between word vectors captures the contextual similarity of words, and is also documented to represent semantic relationships (Mikolov et al., 2013).

We fix the vector space to 200 dimensions ($N = 200$), set the window size to 5 words before or after ($C = 5$), and assume a cosine similarity of 0.4 as the cut-off, following Davis et al. (2020). This is a common setting for applications of the model. When we expand the uncertainty word set, we end up with 53 stems. The three stems most similar to the two seed stems "uncertain" and "uncertainti" are "unstabl*," "unpredict*," "unsettl*" and "jitter," "uncertainli," "anxieti," respectively. We then manually trimmed the model-generated term set, dropping terms that would most likely suggest other notions. Eighteen stems, as synonyms for "uncertain," are selected in the final stage. The top three analogous stems to "UK" are "Britain," "British" and "Countri". We incorporate the first two into the set that refers to the country. A full list of selected terms is provided in Appendix A.

News containing the trio of term sets are collected and this gives us a total of 114,525 articles (after removing duplicates). The monthly frequency of those news articles contributes to our aggregate BUI. As there is variation in the volume of news over time, we scale the raw frequency by the total number of articles in the eleven newspapers over the same period. Following Azzimonti (2018), we take the number of articles containing the word "today" as a proxy for the total number of news items. The index is then normalized to a maximum value of 100.

[8] See Mikolov et al. (2013) for more details.

2.3 Measuring Topic-level BUIs: An Application of a LDA Model

Latent Dirichlet Allocation (LDA) model is an unsupervised topic modelling ML algorithm and is one of the most popular models used in textual analysis. More broadly, supervised and unsupervised learning are the two main tools in ML. Supervised learning refers to the process of inferring a function that maps observations to labels, i.e., classes, by learning the observation-label pairs. With this function, the classes for unlabeled observations can be predicted. In this respect, pre-determined search words are not required, and all the words in the corpus can be taken into consideration. Yet, in our case, using these algorithms requires a substantial number of manually labeled content-topic pairs, making it highly labour-intensive, especially considering the variability of the English language and lexical usage. Such pitfalls can be avoided by using unsupervised learning models, in which labeled dataset is not required. These models can self-learn the underlying structure of the data and look for hidden patterns, i.e., classes within the data and can automatically match the patterns to observations. Intuitively, in our setup, these algorithms can self-uncover latent topics in the corpus, and map news content to these topics.

LDA is a Bayesian probabilistic model. Suppose there is a corpus with V unique words and D documents. The first objective of the LDA model is to extract a predefined number of K latent patterns, or so-called "topics," from the corpus, with each topic k being a probability distribution over V unique words from our corpus denoted by vector φ_k. Intuitively, a topic is a grouping of words, each of which contributes differently to that topic. The LDA model also estimates the topic distribution of each article. In this sense, each article is represented by a mixture of topics with different "weights." Technically speaking, each document d can be interpreted as a probability distribution over K topics denoted by vector θ_d.

LDA is a generative statistical model in which the estimation process generates our corpus with two distributions φ_k and θ_d. The probability of the ith word appearing in document d is $p(d_i) = \sum_k \varphi_k^v \theta_d^k$, where φ_k^i is the probability of the ith word appearing in the topic k, and θ_d^k is the probability of topic k in document d. Then, the probability of accurately generating our corpus is $p(C) = \prod_{d=1}^{D} \prod_{i=1}^{V} p(d_i)^{n_{di}}$, where n_{di} represents the number of occurrences of ith word in document d.

The aim in the generative process is therefore to find the parameters for the two distributions that maximize $p(C)$. An important step in the LDA process is to place Dirichlet priors on the two probability vectors, that is, $\varphi_k \sim Dirichlet(\beta)$ and $\theta_d \sim Dirichlet(\alpha)$, where α and β are the hyperparameters that decide the concentration of the two distributions. A low α results in a steep topic distribution for each article, and the model with a low value of β provides a steep word distribution over each topic. We apply a popular Gibbs Sampling estimator for model estimation (Griffiths and Steyvers, 2004).

Essentially, LDA reduces the dimensionality of our corpus, with V dimensions (V unique words) for the original data, and afterwards, K dimensions (K topics). It is worth noting that each topic is the probability distribution over V words. Hence, the LDA accounts for the full V dimension, i.e., all words. Appendix B provides a more detailed description.

As for model selection, we set the number of topics K to 80.[9] In this process, perplexity, a statistical measure of how well the generative model predicts samples, is often used to evaluate the performance of the LDA model. Yet, good statistical indicators may coexist with low interpretability of the output topics. Hence, models should be evaluated on the basis of real-world performance in specific tasks but not the technical criteria (see Chang et al. (2009)). Accordingly, what we value most in the process of parameter selection is the extent to which the topics of interest are reasonably grouped. Fewer topics will result in topics of interest being mixed up, and a greater number of topics can lead to topics that are difficult to interpret.

Qualified news articles, described in Section 2.2, are cleaned according to the process outlined in Section 2.1 and used as training corpus in the LDA model estimation. Before feeding the corpus into the LDA model, it is usual to exclude frequent and rare words from the corpus to obtain a more interpretable model fit, i.e., grouping of topics (Gentzkow et al., 2019). To this end, we filter extreme words by removing word stems that occur in less than 300 documents and more than 70% of all the documents. The result is that we end up with 8,380 unique word stems. The model estimation usually stabilizes within a few hundred iterations. We therefore set the number of iterations in our estimations to 2,000.

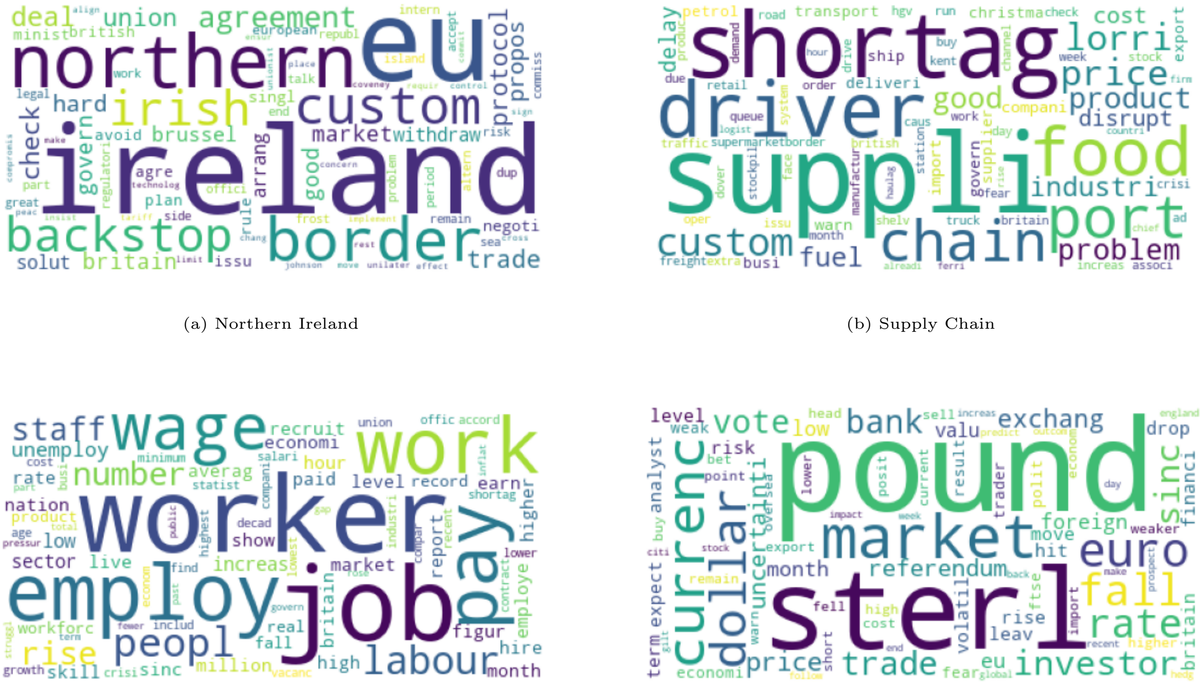
The LDA model generates two distributions of interest, the distribution of words in each topic and the distribution of topics in each article. Notably, the estimated LDA model fixes the first distributions, i.e., topics. The model then serves to infer the topic distribution of the articles, either from the original training corpus or from new, unseen texts. This feature allows for simple and consistent index updates with newly added articles.[10]

Figure 2 provides a visualization of the outputs from the first distribution. Each word cloud represents a topic with the size of the words reflecting their relative weight in the topic. We define each topic based on the most important words contained in that topic. For example, Figure 2 (a), (b), (c) and (d) are example topics (word mixtures), that we identify as Northern Ireland, supply chain, employment, and macroeconomy-currency relevant topics, respectively.[11] Then, based on the second distribution, we compute the share

[9] We rely on the hyperparametric optimization techniques under LDA model estimation using the popular 'MALLET' package to achieve self-tuned α and β that are optimized every ten iterations. Robustness checks for different hyperparameter values are computationally expensive so, in line with other studies, we only perform one set of hyperparameters (Hansen and McMahon, 2016; Hansen et al., 2018; Larsen et al., 2021).

[10] We estimate the LDA model using news articles from our corpus up to November 2021 giving us 111,797 articles in total. If a new compelling topic of Brexit uncertainty arises it would be possible to re-estimate the LDA model.

[11] Word cloud graphs for the remaining 76 topics are available from the authors upon request.



(a) Northern Ireland

(b) Supply Chain

(c) Employment

(d) Macroeconomy-Currency

Figure 2: Example outputs of the LDA model. Each word cloud represents a topic, i.e., a probability distribution of words. The size of a word indicates the probability of the word appearing in the topic. Each word cloud shows only the 80 words with the highest probability of occurrence. The labels of the topics are manually defined.

of each topic in all news published in a given month and subsequently compute the topic frequencies. This gives rise to K monthly time series. i.e., topic-specific BUIs. To keep the topic-level and the aggregate BUIs comparable, the topic-specific indices are scaled in the same way that we scale the aggregate index in the normalization process.[12]

3 Brexit Uncertainty Index

In this section, we present our Brexit uncertainty indices and our analysis of the amplification effect of the COVID pandemic on our BUIs. We present, in turn, the aggregate BUI, the aggregate BUI after removing the effect of COVID, topic-level BUIs and finally, topic-level BUIs excluding COVID effects.

[12] For example, if the maximum value of the aggregate index before normalization is 200, then the normalization of the aggregate BUI is to divide all index values by two in order to scale the index to a maximum value of 100. Subsequently, the topic BUIs are normalized via a division by the number that normalizes the total index to a maximum value of 100 (two in this case).

3.1 Aggregate BUI

Figure 3 presents the estimation of our aggregate BUI. Observe that the index shows obvious spikes around the Brexit referendum date in June 2016, the general election and the start of UK-EU negotiations in June 2017, failed votes on the withdrawal agreement in early 2019, the Brexit extension in late 2019, and the end of the transition period in late 2020. In contrast, the economy experienced lulls in the level of uncertainty following the Brexit deal, i.e., the UK-EU Trade and Cooperation Agreement, that was signed in December 2020. The low and flat curve before 2016 shows that there was little concern about Brexit prior to this date. It is also worth noting that in mid-2020, before the agreement has been reached, there is a clear trough in the index. This can be attributed to the outbreak of COVID-19. As mentioned before, our index is scaled by the total number of news items, i.e., it measures the relative share of Brexit uncertainty in all news. As such, our index rightly falls as attention turned to reporting on the pandemic.

To verify that our index is a reasonable proxy of Brexit related uncertainty, we compare our index with the Brexit uncertainty index generated from the DMP survey. The DMP is a survey that targets business managers and is available from September 2016. In this survey, subjects respond to whether, and to what extent, Brexit is a source of uncertainty for their business. In this case, we compare our BUI with the time evolution of the proportion of managers that rate Brexit as their largest driver of uncertainty.

As shown in Figure 3, there is a clear common movement in the trends of our BUI and the DMP series with a correlation coefficient of 0.82. This correlation climbs to 0.93 after September 2018 (before which the DMP had a relatively small number of respondents, and question about Brexit uncertainty was not asked every month).[13] These high correlations suggest that our index is highly correlated with the reported concerns of UK businesses and gives us confidence that, to some extent at least, we are capturing the degree of Brexit uncertainty at any given point in time.

Note that while the two series fluctuate in roughly the same direction in the post-2020 period, Brexit uncertainty under DMP experiences a relatively sharper decline in early 2020 and remains at a lower level with respect to our BUI. One possible explanation is that our BUI is exaggerated by the COVID pandemic.

3.2 Aggregate BUI Excluding the Effects of COVID

A potential concern is whether the rapid spread of COVID-19 in later 2019 and early 2020 biased our uncertainty indices, given that the pandemic also generated a significant degree of uncertainty for the UK economy and would most likely work in the same direc-

[13] The DMP's membership panel was below 2,000 before July 2017, exceeded 4,000 in May 2018 and is now over 9,000.

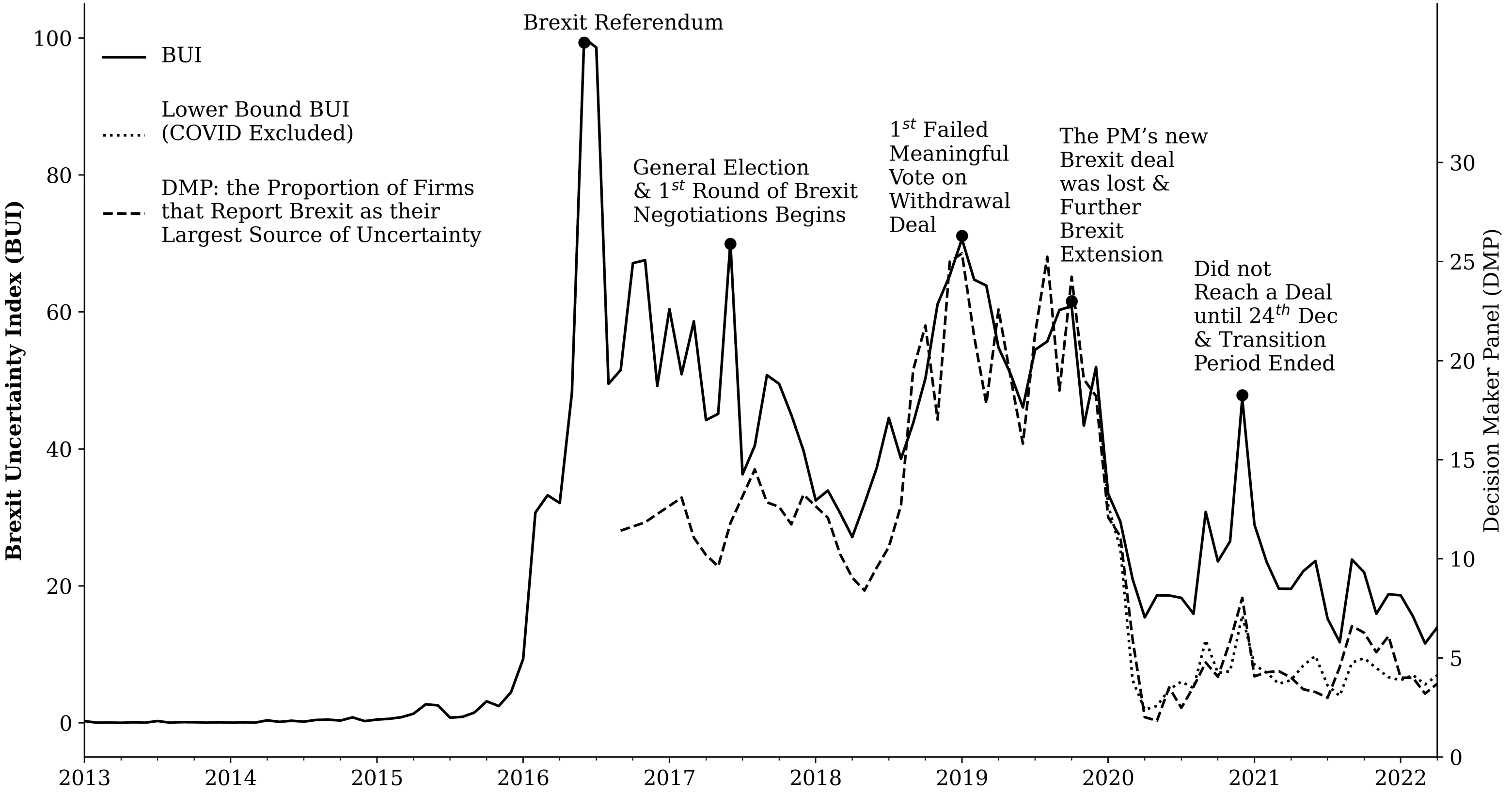


Figure 3: BUI, lower bound BUI (COVID excluded), and the proportion of firms that report Brexit as their largest source of uncertainty from the DMP. BUI is a scaled monthly index based on news articles that include "Brexit" and one or more words in the "uncertainty" word set, and one or more terms in the "UK" term set listed in Section 3.2. The index is normalized to a maximum value of 100. Lower Bound BUI stands for scaled monthly amount of news articles including the identical triplet of terms as BUI but excluding all articles containing one or more words suggesting "COVID" that are issued after January 2020. DMP reflects the proportion of firms that report Brexit as their largest source of uncertainty.

tion as Brexit uncertainty on some issues, for example, supply chain disruptions and labor shortages. In fact, the majority of our eligible post-2020 Brexit uncertainty news items also discuss the coronavirus pandemic (13,420 out of 22,594 articles). It is therefore hard to say to what extent the uncertainty being discussed in the news truly stems from Brexit, rather than the pandemic. Accordingly, any news-based measure of Brexit uncertainty may overestimate Brexit related uncertainties. To capture Brexit uncertainty or lower bound BUI, all news published after 2020 containing COVID-related words have been excluded.[14]

Figure 3 shows the lower bound BUI. Clearly, post-2020, this index and DMP series move together very closely with a correlation coefficient of 0.90 (compared to 0.62 between the aggregate BUI and the DMP series for the same period). The correlation coefficients are 0.86 overall and 0.97 after September 2018. These findings suggest that our lower bound BUI does succeed in disentangling the uncertainty of Brexit from that of the pandemic and captures purely BUI as a lower bound.

3.3 Topic Specific BUIs

Figure 4 plots eight of resulting topic-level BUIs.[15] Note that the topic-level BUI figures have different scales reflecting the importance of that topic of uncertainty. Looking at the different figures we see that the Northern Ireland BUI (Figure 4(a)) had its greatest peak when the "Irish Backstop" was replaced by the new "Northern Ireland Protocol" that guarantees no customs checks or controls on the island of Ireland, i.e., no hard border, but rather a de facto customs border in the Irish Sea. Similarly, the Immigration BUI spiked five times (see Figure 4(b)), first around the Brexit referendum, the second at the commence of UK-EU negotiation, the third on the publishing of post-Brexit immigration system white paper, the forth at the deadline of the EU settlement scheme, and finally during the time when migrants were first seen crossing the English Channel in late 2021.

For other topics, Figure 4(c) shows that the supply chain index peaked at the end of transition period, when uncertainty on reaching a Brexit deal prompted unprecedented supply chain concerns. More recently, in late 2021, post-Brexit Britain is again facing a significant supply chain crisis. Labor shortages, represented by the shortage of truck drivers, has impacted the supply of energy and food. This trend is reflected in the not only supply chain BUI, but also the energy & climate (Figure 4(d)) and the employment (Figure 4(e)) BUIs.

The trade policy BUI (Figure 4(f)) shows a large number of fluctuations since 2016, with significant responses when the leaving date was confirmed in mid-2017, and periods of major concern surrounding the signing of the Brexit deal: the three failed meaningful

[14] COVID related word stems are "COVID," "coronavirus," "pandem," "vaccin," and "epidem."

[15] We only show topic-level BUIs from 2015 onwards, i.e., without the 2013 to 2015 period, as all topic-level indices prior to 2015 are close to zero and barely fluctuate.

votes on the withdrawal deal in early 2019 and in late 2021 when Brexit deal was reached only a week before the end of transition period. The fishing BUI (Figure 4(g)) that mainly captures worries about possible changes to the fishing rights of EU vessels in UK waters and those of UK vessels in EU waters surged during the signing of the Brexit deal that set out fishing licence issues, the May 2021 protests over fishing rights in Jersey, and a series of intensified British-French conflicts over fishing licences in late 2021. The macroeconomy BUI (Figure 4(h)) that covers tax, government spending and the budget, interest rates, inflation rates, currencies, financial markets and economic growth, rose sharply during the Brexit referendum (note the scale of the index is significantly higher than the other seven topi-level BUIs shown in Figure 4).

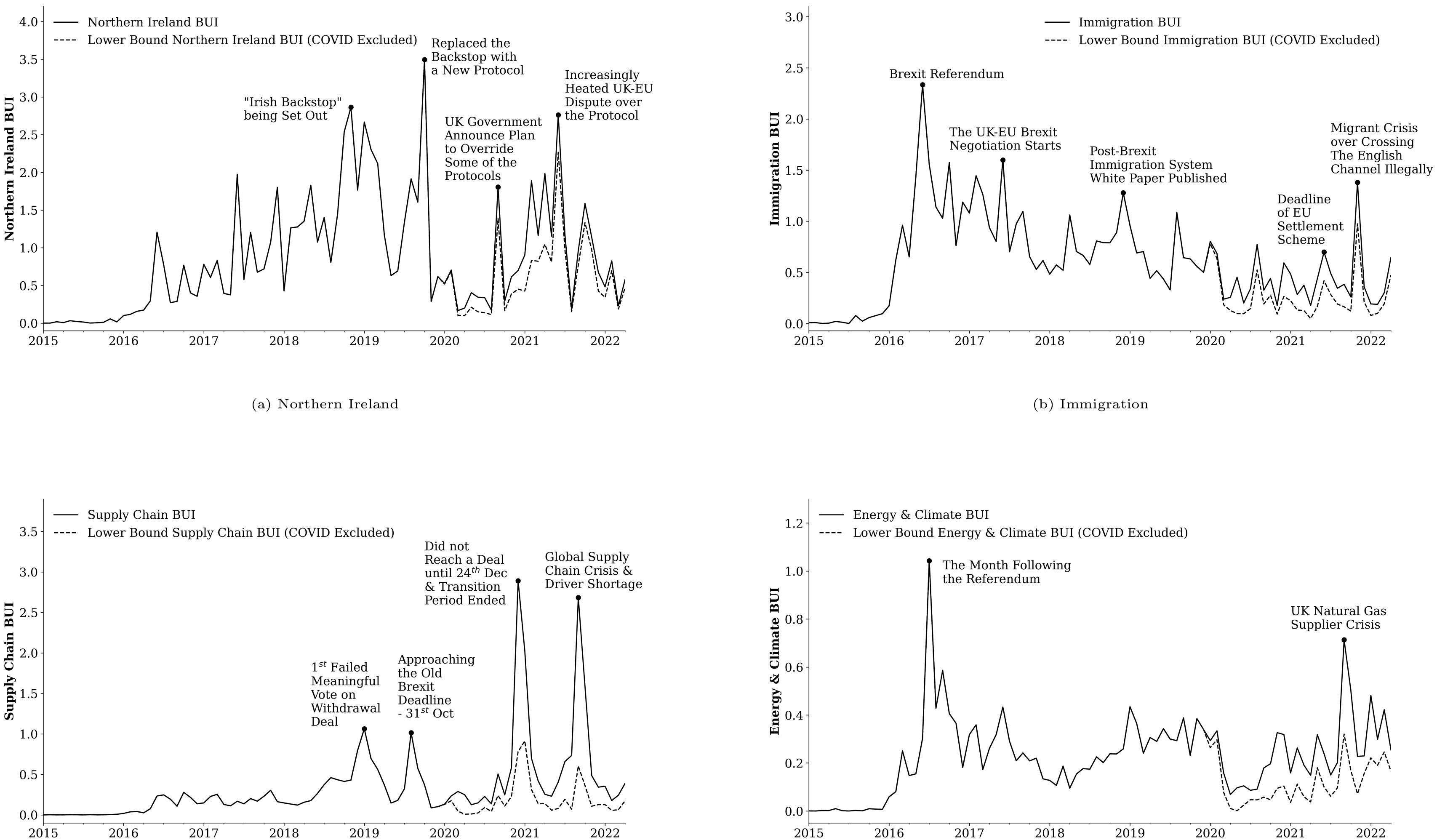
To further probe our topic-level BUIs, we report 14 different BUIs in Table I. Index values are reported for five different periods that capture the four main Brexit events. The five periods are (1) the pre-referendum period, from January 2013 to January 2016, (2) the referendum period, from the announcement of the referendum date in February 2016 to May 2017, (3) the negotiation period, from the start of the UK-EU negotiations in June 2017 to December 2019, (4) the transition period, from January 2020, when the UK formally leaves the EU and enters the transition period, to December 2020, when the Brexit deal has been reached and the transition period ends, and (5) post-Brexit periods, from January 2021 to April 2022.

In general, Northern Ireland and trade policy-related Brexit uncertainty have received the most attention.[16] During the pre-referendum period, Brexit uncertainty was barely evident. Later, the macroeconomy-currency and trade policy topics led to some of the greatest uncertainty in the referendum period. The Northern Ireland issue is the largest source of uncertainty in the negotiation, transition, and post-Brexit periods, as the centre of a dispute between the UK and the EU.[17]

It is also clear that COVID matters as we saw in Figure 4, especially for the employment, macroeconomy-government spending & budget, and supply chain issues. This conclusion can also be drawn from Table I, where the joint effects of the COVID and Brexit on these three issues are three times that of Brexit uncertainty alone. The role of COVID on energy & climate BUI is also apparent (shrinkage to 46.7% after removing the amplification effect of COVID), reflecting the recent energy crisis exacerbated by both events. In contrast, Northern Ireland and the fishing dispute BUIs appear to be the least exposed to the pandemic.

[16] Here, we do not consider the macroeconomy topic as a whole, but discuss its subtopics.

[17] Some topics are more likely to appear together in the same article, with the most frequent topics appearing together being tax and government spending budget (correlation of .10 in articles), immigration and employment (.08), as well as trade policy and food (.06). Details of correlations between topics are shown in Appendix C.

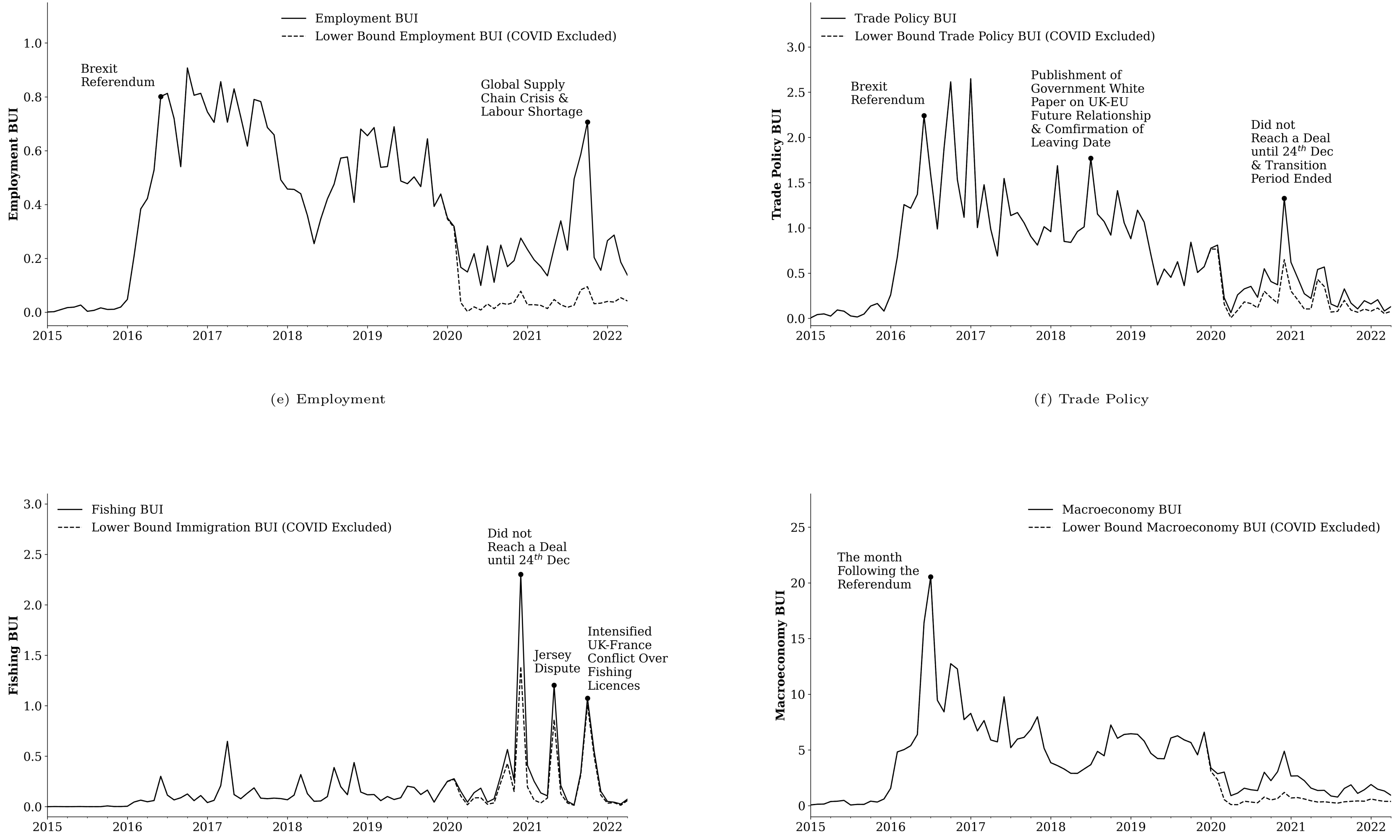


(a) Northern Ireland

(b) Immigration

(c) Supply Chain

(d) Energy & Climate



(e) Employment

(f) Trade Policy

(g) Fishing

(h) Macroeconomy

Figure 4: Topic specific BUIs. Topic specific BUIs indicate scaled monthly share of news coverage concerning each topic in question. The indices are normalized by dividing by the number that scales the maximum value of the aggregate BUI to 100. Lower Bound BUIs stands for the generated BUIs after excluding all articles mentioning COVID issues.

Table I: Brexit Uncertainty by Topic and Time Period, 2013-2022

Time Period	2013:1-2016:1	2016:2-2017:5	2017:6-2019:12	2020:1-2020:12	2021:1-2022:4	2020:1-2022:4	2013:1-2022:04
	Pre Referendum Period	Referendum Period	Negotiation Period	Transition Period (COVID Excluded)	Post-Brexit Period (COVID Excluded)	COVID Period	Overall (COVID Excluded)
Aggregate Brexit Uncertainty	0.95	55.45	48.6	24.9	19.03	21.55	27.07
				(42.23%)	(37.37%)	(39.78%)	
Macroeconomy	0.14	8.96	5.37	2.40	1.57	1.92	3.29
				(35.43%)	(27.90%)	(31.92%)	
Government Spending & Budget	0.01	0.82	0.59	0.34	0.24	0.28	0.35
				(29.03%)	(22.20%)	(25.70%)	
Tax	0.01	0.64	0.43	0.19	0.21	0.20	0.26
				(41.01%)	(32.44%)	(35.94%)	
Currency	0.02	1.46	0.51	0.17	0.06	0.11	0.38
				(37.80%)	(20.58%)	(32.26%)	
Housing Price	0.00	1.02	0.73	0.27	0.15	0.20	0.40
				(43.34%)	(16.22%)	(31.81%)	
Trade Policy	0.03	1.46	0.95	0.48	0.27	0.36	0.57
				(62.99%)	(56.27%)	(60.09%)	
Northern Ireland	0.01	0.49	1.45	0.52	1.11	0.86	0.69
				(70.41%)	(70.76%)	(70.67%)	
Supply Chain	0.00	0.15	0.35	0.48	0.73	0.62	0.27
				(32.72%)	(30.14%)	(31.00%)	
Energy & Climate	0.00	0.34	0.25	0.19	0.30	0.25	0.18
				(47.48%)	(46.33%)	(46.69%)	
Fishing Dispute	0.00	0.14	0.14	0.38	0.29	0.33	0.14
				(67.08%)	(74.75%)	(70.95%)	
Immigration	0.02	1.17	0.74	0.44	0.44	0.44	0.49
				(64.61%)	(55.87%)	(59.63%)	
Employment	0.01	0.67	0.54	0.21	0.29	0.25	0.31
				(37.19%)	(13.72%)	(22.12%)	
Food Industry	0.00	0.31	0.37	0.38	0.29	0.33	0.23
				(68.86%)	(61.38%)	(65.09%)	
Manufacturing	0.01	0.60	0.72	0.30	0.18	0.23	0.34
				(44.10%)	(47.57%)	(45.63%)	

Notes: A summary of the index values for the aggregate BUI and 14 topic-level BUIs of Brexit uncertainty averaged over time. The percentages (in parentheses) are the proportions after removing articles that mention COVID.

4 Evaluation

As we note in Section 3.1, our aggregate BUI is closely related to the index derived form business concerns towards Brexit uncertainty from the DMP survey. This section provides further validation exercises. We begin with a comparison of BUI with an UK economic policy uncertainty (EPU) index constructed by Baker et al. (2016) (henceforth BBD-EPU). Next, we also present a new UK EPU index that uses our ML approach and compares it to the previously calculated BBD-EPU index. As a further robustness check, we examine whether the political leanings of the newspapers included in our study change our results. Finally, we compare how closely our BUIs can be matched to UK stock market volatility.

4.1 Comparison to the BBD-EPU Index

The BBD-EPU index is constructed using the dictionary method, i.e., based on the frequency of news items that contain pre-defined sets of words. As can be seen in Figure 5, the BBD-EPU index and our aggregate BUI have followed very similar trends since 2016 with a correlation of 0.73. Our BUI, nevertheless, reacts more strongly to the post-referendum Brexit event, benchmarked against the EU referendum. In contrast, the BBD-EPU index shows a stronger response to the COVID-19 epidemic in 2020. While the two indexes encapsulate different elements, the uncertainty associated with Brexit is clearly the most important factor affecting UK economic policy in recent years. It is reassuring that the indices follow similar trends.

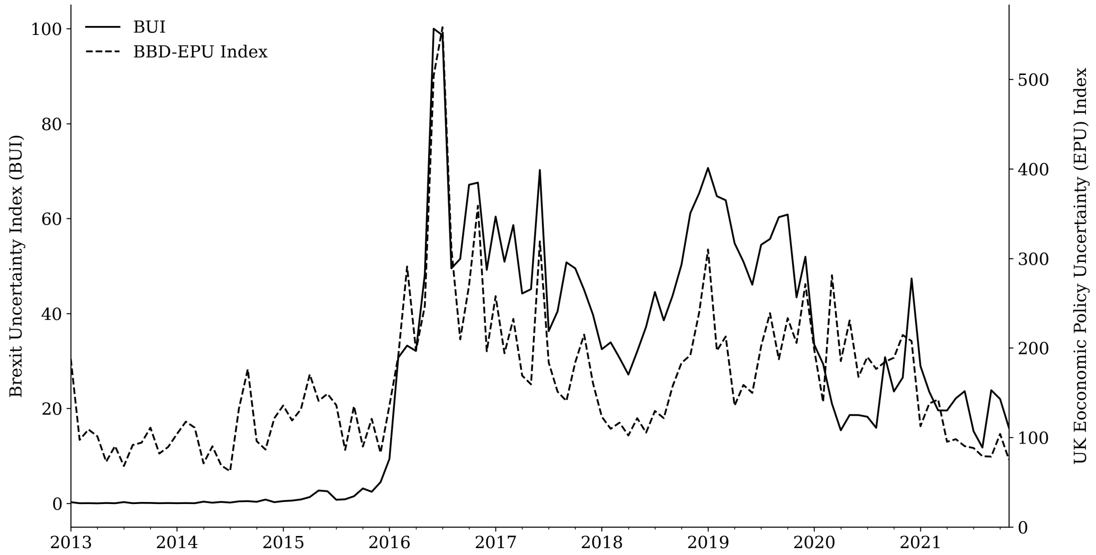


Figure 5: Our BUI and BBD-EPU Index. The figure shows our aggregate BUI from Figure 1 and the BBD-EPU index for the UK economy.

4.2 A new EPU Index using a ML Approach

The BBD-EPU index is based on the frequency of news containing one or more term from the trio sets indicating “Economic”, “Policy” and “Uncertainty”, respectively. We, instead, collect the news articles with their “Economic” term set and our “Uncertainty” term set without regard to “Policy.” We then rely on the LDA model to distinguish economic “policy” relevant content from collected articles. The resulting index, shown in Figure 6, exhibits a clear co-movement with the BBD-EPU index, with a correlation coefficient of 0.83. The correlation coefficient is only 0.65 before the LDA model was used to select economic policy-related content.

The strong correlation between the BBD-EPU index and the EPU index developed with our ML approach reveals three insights. First, the selection of our “uncertainty” term set is reasonable. Second, the LDA model helps to identify policy-related contents. Accordingly, we confirm that the LDA model can help to disentangle the Brexit and its topics from a large number of other elements. Finally, it demonstrates, we believe, that our indices are fairly reliable proxies for different Brexit uncertainties.

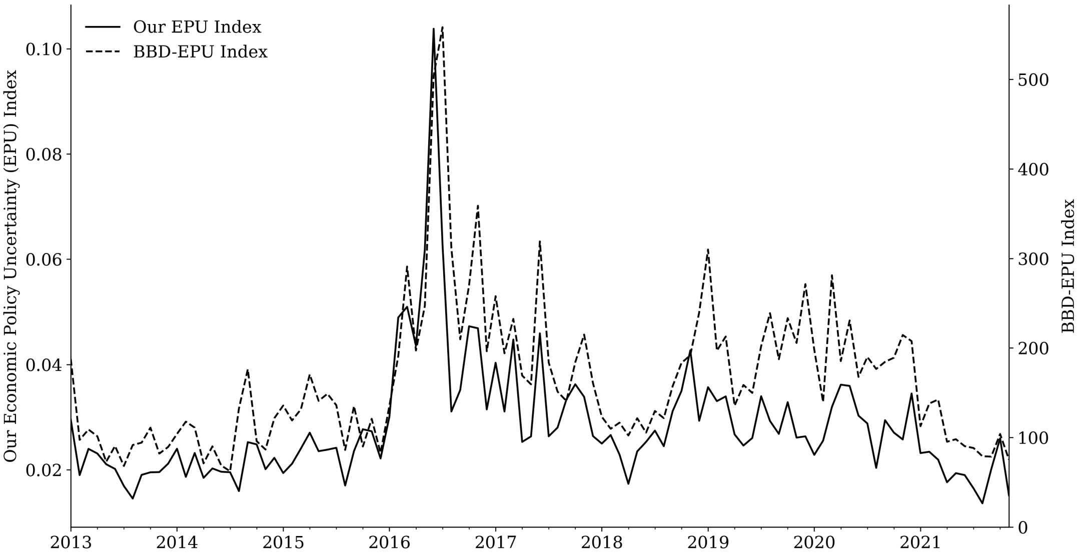


Figure 6: Our EPU Index and the BBD-EPU Index. The figure shows our EPU index constructed with our BUI measure and the BBD-EPU index constructed by Baker et al. (2016). Both EPU indices are based on the same 11 leading UK newspapers listed in Section 2.1.

4.3 Political Leanings in Newspaper Coverage of Brexit Uncertainty

The political leaning of a newspaper has the potential to bias newspaper coverage of Brexit uncertainty. The Conservative Party, typically considered to be on the center-right of the political spectrum, has been in power for the entire period of our analysis. Our measure, then, may be skewed if left (right) leaning newspapers badly overstate (downplay) the extent of Brexit uncertainty. To this end, we examine how the political leanings of newspapers affect our BUI. The results show that the BUIs generated from left-wing and right-wing newspapers move together closely, with a correlation coefficient of 0.98.[18] This means that the political slant of newspaper coverage has no impact on our BUI measures.[19]

4.4 Comparison to Stock Market Volatility

Finally, another way to evaluate the usefulness of our BUIs is through a comparison with other uncertainty indicators of the UK economy. One obvious comparator is the FTSE 100 Volatility Index (VFTSE), representing the implied volatility on the FTSE 100, available until June 2019.[20] When we compare the two we observe that our aggregate BUI and the currency BUI are the two that most closely match the VFTSE. As shown in Figure 7, our aggregate BUI and VFTSE have followed similar trends after the Brexit referendum (a correlation of 0.60 over this period). Since the referendum, our currency BUI and VFTSE follow similar trends with a correlation of 0.58 which is intuitive since the strength of the currency has long been considered an important factor affecting the stock market. However, after 2018 the VFTSE reacts to the surge in U.S. stock market volatility in early 2018, and other international influences, e.g., the US-China trade war in late 2018, yet, our BUI does not react to such events.[21]

An alternative comparator is the FTSE 100 realized volatility, available for the whole time span of our index.[22] We select the pre-2020 period for comparison because after 2020, the COVID shock had a significant impact on the stock market that overshadowed Brexit uncertainty effects and is a period when stock market volatility and Brexit uncertainty are not significantly correlated. Our currency and aggregate BUIs are correlated to the FTSE 100 realized volatility after the referendum (correlations of 0.68 and 0.54, respectively).

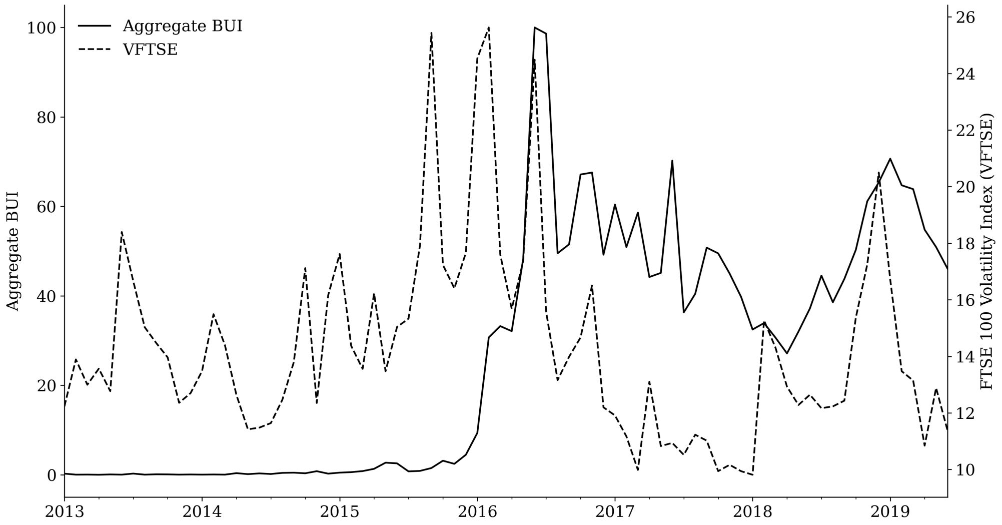
[18] Left-wing newspapers include *The Guardian* and *The Mirror.* Right-wing newspapers include *The Daily Telegraph, The Times, The Daily Mail and Mail on Sunday* and *The Daily Express.* The source of the political leanings of newspapers: https://www.oxford-royale.com/articles/a-guide-to-british-newspapers/. The correlation is 0.97 after removing the centre-left or right newspaper, *The Times.*

[19] Figures showing the index values by political leaning are available from the authors upon request.

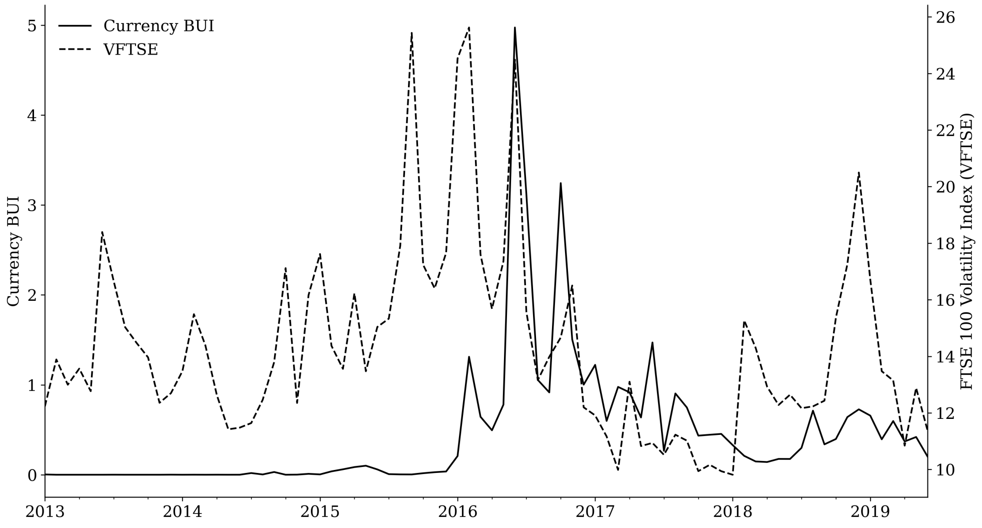
[20] Datasource: Bloomberg.

[21] The VIX, i.e., the implied volatility of the U.S. stock market, spiked in early 2018.

[22] Datasource: "Oxford-Man Institute's realized library" (Heber et al., 2009).



(a) Aggregate BUI Compared to FTSE 100 Volatility Index (VFTSE) from Jan 2013 to Dec 2019



(b) Currency BUI Compared to FTSE 100 Volatility Index (VFTSE) from Jan 2013 to June 2019.

Figure 7: Aggregate BUI, Currency BUI, and FTSE 100 Volatility Index (VFTSE)

5 Conclusions

The uncertainties triggered by the withdrawal of the UK from the EU are large and protracted. From the pre-referendum era to the current post-Brexit period the risks associated with the withdrawal have generated various uncertainties including, but not limited to, supply chains, employment, immigration, and the macroeconomy more generally. A major concern is that these uncertainties affect the economic environment in the UK, for example, investment, trade, and employment. However, quantifying the degree of Brexit uncertainty is a challenge.

In this paper we adapt two ML techniques from computational linguistics to develop a novel news-based measure of aggregate and topic specific Brexit uncertainty faced by the UK. Our methodology allows us to consider all news pertaining to Brexit uncertainty and decompose the news by topic to obtain the time-series share of aggregate and topic-level uncertainty. In this way, we can quickly and cost effectively measure the evolution of this uncertainty and its component topics over time and in real-time. We have developed an online tool that enables us to extend the analysis of this paper quickly and efficiently that allows us to keep policymakers and newspapers informed on how uncertainty is evolving.

Validation exercises demonstrate that our indices are reasonable proxies for Brexit uncertainties. We find that our aggregate index is strongly correlated with both the time-series proportion of firms reporting Brexit as the largest source of uncertainty in the DMP survey and the BBD-EPU index developed by Baker et al. (2016). Moreover, our UK EPU index constructed using our ML approaches shows a clear co-movements with the BBD-EPU index. Finally, our currency-related BUI and aggregate BUI follow fairly similar trends to indices of UK stock market volatility between the Brexit referendum and the COVID outbreak.

We also measure the Brexit uncertainty excluding the impact of the COVID pandemic. As expected, this lower-bound index is more closely related to the attitudes of businesses toward Brexit uncertainty during the pandemic period, compared to our aggregate index.

Our measure of Brexit uncertainty opens up a number of possible channels for future research. Specifically, the ability to measure Brexit uncertainty should allow researches to quantify the impact of Brexit in different dimensions. Our approach may also be valuable for policy makers who want to see how a particular event or policy affects different types of Brexit uncertainty or economic policy uncertainty more generally.

References

Ahmad, S., Limão, N., Oliver, S. and Shikher, S. (2020), Brexit uncertainty and its (dis) service effects, Technical report, National Bureau of Economic Research.

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Leather, J., Meyer, B., Mihaylov, E. et al. (2020), 'Economic uncertainty before and during the covid-19 pandemic', *Journal of Public Economics* **191**, 104274.

Arellano, C., Bai, Y. and Kehoe, P. J. (2019), 'Financial frictions and fluctuations in volatility', *Journal of Political Economy* **127**(5), 2049–2103.

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

Azzimonti, M. (2018), 'Partisan conflict and private investment', *Journal of Monetary Economics* **93**, 114–131.

Bachmann, R., Elstner, S. and Sims, E. R. (2013), 'Uncertainty and economic activity: Evidence from business survey data', *American Economic Journal: Macroeconomics* **5**(2), 217–49.

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

Bank of England (2019), 'Monetary policy report - november 2019: In focus - uncertainty and brexit'. available at `https://www.bankofengland.co.uk/monetary-policy-report/2019/november-2019/in-focus-uncertainty-and-brexit`.

Basu, S. and Bundick, B. (2017), 'Uncertainty shocks in a model of effective demand', *Econometrica* **85**(3), 937–958.

Bernanke, B. S. (1983), 'Irreversibility, uncertainty, and cyclical investment', *The quarterly journal of economics* **98**(1), 85–106.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Bloom, N. (2009), 'The impact of uncertainty shocks', *econometrica* **77**(3), 623–685.

Bloom, N., Bunn, P., Chen, S., Mizen, P., Smietanka, P. and Thwaites, G. (2019), The impact of brexit on uk firms, Technical report, National Bureau of Economic Research.

Bloom, N., Floetotto, M., Jaimovich, N., Saporta-Eksten, I. and Terry, S. J. (2018), 'Really uncertain business cycles', *Econometrica* **86**(3), 1031–1065.

Burn, I., Button, P., Corella, L. F. M. and Neumark, D. (2019), Older workers need not apply? ageist language in job ads and age discrimination in hiring, Technical report, National Bureau of Economic Research.

Bybee, L., Kelly, B. T., Manela, A. and Xiu, D. (2020), The structure of economic news, Technical report, National Bureau of Economic Research.

Caldara, D. and Iacoviello, M. (2022), 'Measuring geopolitical risk', *American Economic Review* **112**(4), 1194–1225.

Caldara, D., Iacoviello, M., Molligo, P., Prestipino, A. and Raffo, A. (2020), 'The economic effects of trade policy uncertainty', *Journal of Monetary Economics* **109**, 38–59.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. and Blei, D. M. (2009), Reading tea leaves: How humans interpret topic models, Technical report, Advances in neural information processing systems.

Crowley, M., Exton, O. and Han, L. (2018), Renegotiation of trade agreements and firm exporting decisions: evidence from the impact of brexit on uk exports, Technical report, Society of International Economic Law (SIEL), Sixth Biennial Global Conference.

Davis, S. J., Hansen, S. and Seminario-Amez, C. (2020), Firm-level risk exposures and stock returns in the wake of covid-19, Technical report, National Bureau of Economic Research.

Douch, M., Du, J. and Vanino, E. (2020), 'Defying gravity? policy uncertainty, trade destruction and diversion', *Lloyds Banking Group Centre for Business Prosperity, Research Paper No* **3**.

Fajgelbaum, P. D., Schaal, E. and Taschereau-Dumouchel, M. (2017), 'Uncertainty traps', *The Quarterly Journal of Economics* **132**(4), 1641–1692.

Fernández-Villaverde, J., Guerrón-Quintana, P., Rubio-Ramírez, J. F. and Uribe, M. (2011), 'Risk matters: The real effects of volatility shocks', *American Economic Review* **101**(6), 2530–61.

Gentzkow, M., Kelly, B. and Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**(3), 535–74.

Gilchrist, S., Sim, J. W. and Zakrajšek, E. (2014), Uncertainty, financial frictions, and investment dynamics, Technical report, National Bureau of Economic Research.

Graziano, A. G., Handley, K. and Limão, N. (2020), Brexit uncertainty: trade externalities beyond europe, *in* 'AEA Papers and Proceedings', Vol. 110, pp. 552–56.

Graziano, A. G., Handley, K. and Limão, N. (2021), 'Brexit uncertainty and trade disintegration', *The Economic Journal* **131**(635), 1150–1185.

Griffiths, T. L. and Steyvers, M. (2004), 'Finding scientific topics', *Proceedings of the National academy of Sciences* **101**(suppl 1), 5228–5235.

Hansen, S. and McMahon, M. (2016), 'Shocking language: Understanding the macroeconomic effects of central bank communication', *Journal of International Economics* **99**, S114–S133.

Hansen, S., McMahon, M. and Prat, A. (2018), 'Transparency and deliberation within the fomc: a computational linguistics approach', *The Quarterly Journal of Economics* **133**(2), 801–870.

Hansen, S., McMahon, M. and Tong, M. (2019), 'The long-run information effect of central bank communication', *Journal of Monetary Economics* **108**, 185–202.

Hassan, T. A., Hollander, S., Van Lent, L. and Tahoun, A. (2019), 'Firm-level political risk: Measurement and effects', *The Quarterly Journal of Economics* **134**(4), 2135–2202.

Heber, G., Lunde, A., Shephard, N. and Sheppard, K. (2009), 'Oxford-man institute's realized library'. Oxford-Man Institute, University of Oxford. Library Version: 0.3.

Heinrich, G. (2005), Parameter estimation for text analysis, Technical report, Technical report.

Husted, L., Rogers, J. and Sun, B. (2020), 'Monetary policy uncertainty', *Journal of Monetary Economics* **115**, 20–36.

Javorcik, B., Stapleton, K., Kett, B. and O'Kane, L. (2020), Unravelling deep integration: Local labour market effects of the brexit vote, Technical report, CEPR Discussion Paper 14222.

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

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G. and Kapadia, S. (2022), 'Making text count: Economic forecasting using newspaper text', *Journal of Applied Econometrics* (forthcoming).

Larsen, V. H. (2021), 'Components of uncertainty', *International Economic Review* **62**(2), 769–788.

Larsen, V. H. and Thorsrud, L. A. (2019), 'The value of news for economic developments', *Journal of Econometrics* **210**(1), 203–218.

Larsen, V. H., Thorsrud, L. A. and Zhulanova, J. (2021), 'News-driven inflation expectations and information rigidities', *Journal of Monetary Economics* **117**, 507–520.

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

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), ‘Efficient estimation of word representations in vector space’, *arXiv preprint arXiv:1301.3781* .

Appendix

A Word Sets

Table A1 shows the words we use to filter out articles related to Brexit uncertainty from the mass of news. An article containing "Brexit" and at least one word from the "uncertainty" and "UK" sets is taken as Brexit uncertainty related articles. All words or word roots associated with "uncertainty" and "UK" are selected based on the outputs of the Word2Vec model.

Table A1: Word Sets

Brexit	Brexit
Uncertainty	uncertain*; instab*; unstabl*; risk*; unpredict*; volatile*; unclear*; worry*; fear*; tension*; anxiety*; nervous*; jitter*; unsettl*; precar*; unknow*; indecis*; angst*
UK	UK (United Kingdom); British; Britain

B LDA: Model and Estimation

We discard the notation defined in the previous section. Suppose the corpus contains M documents in total, where N_m denotes the total number of words in document m. Let $z_{m,n}$ be the topic for the n-th word in document m, $w_{m,n}$ be the n-th word in document m, K be the number of latent topics, and V be the set of all unique words t in the corpus. The objects of the model are twofold: (a) to estimate the mixture component for each topic, $\Phi = \{\varphi_k\}_{k=1}^{K}$ ($K \times V$ matrix), in the form of probability distributions over V words; (b) to estimate the topic mixture proportion for each document, $\Theta = \{\theta_m\}_{m=1}^{M}$ ($M \times K$matrix), in the form of probability distributions over K topics. α and β are the pre-defined hyperparameters, which determines the prior weight of each topic in a document and prior weight of each word in a topic. In this section, we briefly describe the LDA model and the estimation method. For more detailed information and explanation, see Griffiths and Steyvers (2004) and Heinrich (2005).

B.1 Model

As we mentioned in the previous section, the LDA model is a generative model: LDA generates all the observable words $w_{m,n}$, which in turn generate the whole corpus. With the notation we defined above, the probability that the word $w_{m,n}$ is the term t can be expressed as:

$$p(w_{m,n} = t|\theta_m, \Phi; \alpha, \beta) = \sum_{k=1}^{K} p(w_{m,n} = t|\varphi_k)p(z_{m,n=k|\theta_m})$$

Then, the generation process of the corpus $W = \{w_m\}_{m=1}^{M}$ are

$$p(W|\Theta, \Phi; \alpha, \beta) = \prod_{m=1}^{M} p(w_m|\theta_m, \Phi; \alpha, \beta) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\theta_m, \Phi; \alpha, \beta)$$

B.2 Estimation

We apply the Gibbs sampling algorithm introduced by Griffiths and Steyvers (2004) to estimate the LDA model. Both target distributions Φ and Θ can be interpreted with the observed words $w_{m,n}$ and their corresponding topics $z_{m,n}$. The goal of the inference could therefore be the distribution

$$p(Z|W; \alpha, \beta) = \frac{p(Z|W; \alpha, \beta)}{p(W; \alpha, \beta)}$$

Gibbs sampling algorithms use the full conditional $p(z_i|Z_{-i}, W; \alpha, \beta)$ to simulate this distribution.To generate the full conditional, we first draft joint distribution, that is

$$p(W, Z|\alpha, \beta) = p(W|Z, \beta)p(Z|\alpha)$$

The transition from the left to the right side of the equation relies on conditional independence. These two components can be processed separately, and the joint distribution can be written as:

$$p(Z, W|\alpha, \beta) = \prod_{z=1}^{K} \frac{\Delta(N_z + \beta)}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(N_m + \alpha)}{\Delta(\alpha)}$$

where $N_z = \{n_z^{(t)}\}_{t=1}^{V}$, and $N_m = \{n_m^{(k)}\}_{k=1}^{K}$. Then, employing chain rule and letting word index be $i = (m, n)$, the full conditional distribution can be derived as:

$$\begin{aligned}
p(z_i = k|Z_{-i}, W) &= \frac{p(W, Z)}{p(W, Z_{-i})} = \frac{p(W|Z)}{p(W_{-i}|Z_{-i})p(w_i)} \cdot \frac{p(Z)}{p(Z_{-i})} \\
&\propto \frac{\Delta(N_z + \beta)}{\Delta(N_{z,-i} + \beta)} \cdot \frac{\Delta(N_m + \alpha)}{\Delta(N_{m,-i} + \alpha)} \\
&= \frac{\Gamma(n_k^{(t)} + \beta)\Gamma(\sum_{t=1}^V n_{k,-i}^{(t)} + \beta)}{\Gamma(n_{k,-i}^{(t)} + \beta)\Gamma(\sum_{t=1}^V n_k^{(t)} + \beta)} \cdot \frac{\Gamma(n_m^{(k)} + \alpha)\Gamma(\sum_{k=1}^K n_{m,-i}^{(k)} + \alpha)}{\Gamma(n_{m,-i}^{(k)} + \alpha)\Gamma(\sum_{k=1}^K n_m^{(k)} + \alpha)} \\
&= \frac{n_{k,-i}^{(t)} + \beta}{\sum_{t=1}^V n_{k,-i}^{(t)} + \beta} \cdot \frac{n_{m,-i}^{(k)} + \alpha}{[\sum_{k=1}^K n_m^k + \alpha] - 1} \\
&\propto \frac{n_{k,-i}^{(t)} + \beta}{\sum_{t=1}^V n_{k,-i}^{(t)} + \beta}(n_{m,-i}^{(k)} + \alpha)
\end{aligned}$$

Finally, the two target distributions can be estimated as:

$$\hat{\theta}_{m,k} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^K n_m^{(k)} + \alpha}$$

$$\hat{\varphi}_{k,t} = \frac{n_k^{(t)} + \beta}{\sum_{t=1}^V n_k^{(t)} + \beta}$$

where $n_m^{(k)}$ denotes the number of times that topic k has been observed with a word in document m, and $n_k^{(t)}$ refers to the number of times that term t has been observed with topic k.

C Correlation of Topics in Articles

Which topics appear together more frequently in the same article? To answer this question, we calculate the correlations of topics in the topic distribution of the articles and present in table C1.

Table C1: Correlation of Topics in Articles

Topic	**Associated topics in the same article**
Government Spending & Budget	Tax (0.106)
Tax	Government Spending & Budget (0.106), Housing Price (0.029)
Currency	N/A
Trade Policy	Food (0.058), Supply Chain (0.022)
Housing Price	Tax (0.029)
Northern Ireland	Supply Chain (0.021)
Supply Chain	Food (0.050), Employment (0.027), Trade Policy (0.022)
Energy & Climate	Food (0.047), Manufacturing (0.020)
Immigration	Employment (0.079)
Employment	Immigration (0.079), Supply Chain (0.027)
Food Industry	Trade Policy (0.058), Supply Chain (0.050), Energy & Climate (0.047)
Manufacturing	Energy & Climate (0.020)
Scotland	N/A

Notes: This table shows the top three other associated topics with a correlation of 0.02 or higher for each topic. Listing less than three topics indicates that no more than three associated topics meet the criteria; N/A indicates that no associated topic meets the criteria. Correlation coefficients are in parentheses.