Advances in Data Analysis and Classification (2025) 19:769–793
https://doi.org/10.1007/s11634-024-00596-4

REGULAR ARTICLE

Natural language processing and financial markets: semi-supervised modelling of coronavirus and economic news

Carlos Moreno-Pérez[2] · **Marco Minozzo**[1]

Received: 18 October 2022 / Revised: 23 August 2023 / Accepted: 1 November 2023 /
Published online: 19 June 2024

Abstract

This paper investigates the reactions of US financial markets to press news from January 2019 to 1 May 2020. To this end, we deduce the content and uncertainty of the news by developing apposite indices from the headlines and snippets of The New York Times, using unsupervised machine learning techniques. In particular, we use Latent Dirichlet Allocation to infer the content (topics) of the articles, and Word Embedding (implemented with the Skip-gram model) and K-Means to measure their uncertainty. In this way, we arrive at the definition of a set of daily topic-specific uncertainty indices. These indices are then used to find explanations for the behavior of the US financial markets by implementing a batch of EGARCH models. In substance, we find that two topic-specific uncertainty indices, one related to COVID-19 news and the other to trade war news, explain the bulk of the movements in the financial markets from the beginning of 2019 to end-April 2020. Moreover, we find that the topic-specific uncertainty index related to the economy and the Federal Reserve is positively related to the financial markets, meaning that our index is able to capture the actions of the Federal Reserve during periods of uncertainty.

Keywords COVID-19 · EGARCH · Latent Dirichlet Allocation · Investor attention · Uncertainty indices · Word Embedding

The views expressed in this paper are the authors' and do not necessarily reflect those of the Bank of Spain or the Eurosystem.

✉ Marco Minozzo
marco.minozzo@univr.it

Carlos Moreno-Pérez
carlos.moreno@bde.es

[1] Present Address: Department of Economics, University of Verona, Verona, Italy

[2] Present Address: Directorate General Economics, Statistics and Research, Bank of Spain, Madrid, Spain

Mathemamatics Subject Classification 91-05 · 91-08 · 68T07 · 68T50

JEL Classification C45 · C58 · D81 · G15

1 Introduction

During 2019, US financial markets rose steadily despite the growing concern about a possible trade war between the US and China, and a no-deal Brexit. At the beginning of 2020, in particular on 19 February 2020, the S&P 500 index reached a historic peak. Then, the spread of COVID-19 in European countries and in Asia led to a memorable collapse of the financial markets, followed by a quick recovery due to the interventions of the Fed and of the US government's fiscal packages. In this paper, we investigate the relation between newspaper articles and financial indices, from the beginning of 2019 until mid 2020, using unsupervised machine learning techniques for text mining.

In the economic literature, text mining techniques are becoming increasingly popular to investigate the effect of the news on the real economy and on the markets. For example, Kalamara et al. (2022) make extensive use of text mining techniques for extracting information from three leading UK newspapers, to forecast macroeconomic variables with machine learning methods. Hansen and McMahon (2016) use unsupervised machine learning methods, in particular Latent Dirichlet Allocation (LDA), for constructing text measures of the information released by the Federal Open Market Committee (FOMC), to investigate the impact of FOMC communications on the markets and on some economic variables. Similarly, Hansen et al. (2018) use LDA and dictionary methods to study the effect of transparency on the decisions of the FOMC. Other papers also investigate the communications of the FOMC using LDA, such as Edison and Carcel (2021), and Jegadeesh and Wu (2017).

Machine learning techniques are also used to build measures of uncertainty based on various text sources. For instance, Ardizzi et al. (2019) construct Economic Policy Uncertainty (EPU) indices for Italy from newspaper and Twitter data to study debit card expenditure. In particular, Soto (2021) uses unsupervised machine learning techniques to construct uncertainty measures from the text information released by commercial banks in their quarterly conference calls. He uses the Skip-gram model for Word Embedding and K-Means to find the word vectors nearest to the vector representations of the words 'uncertainty' and 'uncertain' and thereby constructs a list of uncertainty words, whose frequency in the documents is used to build an uncertainty index. Then, with the help of LDA, he constructs topic-specific uncertainty indices. On the other hand, an example of derivation of uncertainty measures from newspaper articles is given by Azqueta-Gavaldón et al. (2023). These authors use Word Embedding (with the Skip-gram model) and LDA to construct national uncertainty indices from Italian, Spanish, German, and French newspapers. Then, they use a Structural VAR model to investigate the impact of the national uncertainty indices on some macroeconomic variables such as investment in machinery and equipment. Other authors also investigate the use of sentiment indices based on various text sources concerning news on the financial markets. Just to mention, Zhu et al. (2019) utilize a monthly text volatility index named the Equity Market Volatility (EMV) and the daily VIX index

to predict the evolution of US financial markets. In particular, they use a GARCH-MIDAS model to incorporate variables with different frequencies (daily and monthly) and conclude that the EMV index is more helpful than the VIX index in predicting volatility.

As far as the COVID-19 pandemic is concerned, Baker et al. (2020) construct three measures to capture different sources of uncertainty: stock market volatility, EPU, and unsureness in business expectations. On the other hand, Haroon and Rizvi (2020) investigate how sentiment has driven financial markets during the first months of the coronavirus pandemic. These authors use an EGARCH model to study the effect of sentiment and panic in investors (using the Ravenpack Panic Index and the Global Sentiment Index) on the volatility of a wide range of financial indices relative to the world and US markets and to 23 sectors of the Dow Jones. In a similar fashion, Albulescu (2020) investigates the effect on the VIX index of the US EPU index, the number of COVID-19 cases, and the COVID-19 death rates. They find that the Chinese and world COVID-19 death rates are positively associated with the VIX index and that the US EPU index is positively associated with the volatility in the financial markets. Moreover, to deepen the analysis, a few authors also proceeded to create their own sentiment indices.

In this paper, we create text measures to quantify the content and uncertainty of US news, related in particular to the COVID-19 pandemic, using unsupervised machine learning algorithms such as LDA, Word Embedding (with the Skip-gram model), and K-Means. In particular, we construct text measures from the headlines and snippets of articles in the English version of The New York Times from 2 January 2019 to 1 May 2020. We concentrated on this time period since it covers the outbreak and the first months of the pandemic of COVID-19, which greatly affected the financial markets and the economy and made it one of the most relevant periods in recent decades, with the aim to evaluate the capabilities of some machine learning techniques. The choice of The New York Times is due both to the availability to the researchers of an electronic database and to the fact that it has a large public in the US and worldwide, whose news is also echoed in other newspapers and media. To infer the content or theme of the news in the documents, that is, in the newspaper articles, we run LDA with sixty topics. Then, we determine the daily probability distribution of each topic and use it as a daily measure of attention to each topic in the daily news. To create uncertainty measures, we resort to Word Embedding (using the Skip-gram model) and K-Means. With these, we come out with a list of words having a meaning similar to the word ‘uncertainty’. Actually, we consider in this list all the words that are in the same clusters of the words ‘uncertain’, ‘uncertainty’, ‘fears’, ‘fears’, and ‘worries’, since they share a similar semantic meaning. This list is then used as an uncertainty dictionary to construct a daily uncertainty index by counting the frequency of its words present in all the articles of a given day. To create topic-specific uncertainty indices, we then combine the daily LDA probabilities of each topic with the uncertainty index obtained with Word Embedding and K-Means. In this way, we come out with uncertainty indices for specific topics such as, in particular, ‘coronavirus’, ‘trade war’, ‘climate change’, ‘economic-Fed’, and ‘Brexit’. To the best of our knowledge, this is one of the first papers to use LDA and Word Embedding to construct topic-specific uncertainty indices for ‘coronavirus’ and ‘trade war’ news, covering, in particular, the first wave

of the COVID-19 pandemic. A similar work has been done by Mamaysky (2023) who built several topic-specific sentiment indices for coronavirus news. He selected news mentioning the words 'coronavirus' and 'COVID-19', from the beginning of 2019 to the end of April 2020, and then applied LDA to classify these news under nine headings. In such a manner, he constructed a daily positive–negative sentiment index with the Loughran-McDonald dictionary (Loughran and Mcdonald 2011) and created topic-specific positive–negative sentiment indices to investigate how they are correlated with the evolution of the stock markets.

In this work, we concentrate on LDA and Word Embedding since they are among the most known and used unsupervised machine learning techniques for text analysis. According to Hansen et al. (2018), these techniques have significant advantages over keywords and dictionary methods since they use all the terms in the corpus to depict paragraphs in a low-dimensional space, in lieu of using parts of them, and identify the most significant words in the data rather than imposing them. In addition, LDA and Word Embedding, being unsupervised methods, have an advantage over supervised methods, such as the FinBert model by Huang et al. (2023), which uses a sample of researcher-labeled phrases from analyst reports, since they do not require a preliminary manual classification of the text to obtain a suitable training set.

In the last part of the paper, we then investigate, implementing some EGARCH models, the relationship between these topic-specific uncertainty indices and the returns of several US financial indices such as the S&P 500, the Nasdaq, and the Dow Jones, as well as the 10 year US treasury bond yields. We find that in the period under scrutiny, the 'trade war' and 'coronavirus' uncertainty indices have a significant negative effect on the mean returns of the S&P 500. In particular, the 'trade war' uncertainty index accounts for most of the behavior of the S&P 500 during 2019, whereas the 'coronavirus' uncertainty index accounts for most of the behavior of the S&P 500 in the first months of 2020. Moreover, an increase in the 'trade war' and 'coronavirus' uncertainty indices significantly increases the volatility of the S&P 500 returns and the mean returns of the VIX index. Our findings on the 'trade war' uncertainty index are in line with those of Burggraf et al. (2020), which suggest that tweets from US President Donald Trump's Twitter account related to the trade war between the US and China had a positive effect on the VIX index and a negative effect on the S&P 500 returns. On the other hand, our findings on the effects of the 'coronavirus' uncertainty index on the financial markets are similar to those of Baker et al. (2020) and Haroon and Rizvi (2020), which investigated the association between the panic for the coronavirus crisis at the beginning of 2020 and the increase in volatility in the financial markets. Besides, we also find that a rise in the 'economic-Fed' uncertainty index significantly increases the mean returns of the S&P 500 index. This would mean that news about interventions of the Fed or the US government has a positive effect on the S&P 500 in days of uncertainty. For instance, this latter index catches the reduction of interest rate by the Fed on the third of March 2020, the 'Emergency Lending Programs' deployed by the Fed on the 17*th* of March 2020, as well as the discussion of the Trump's fiscal package.

The paper is organized as follows. In Sect. 2 we introduce our text data and explain the construction of the topic-specific uncertainty indices with the help of LDA, Word

Embedding, and K-Means. In Sect. 3 we illustrate the EGARCH analysis and comment on the results. Finally, in Sect. 4 we give some conclusions.

2 Topic and uncertainty analysis of newspaper text data

2.1 The New York Times data

Our raw data are the headlines and the snippets of the English version of the articles of The New York Times from 2 January 2019 to 1 May 2020. The snippets are small pieces of information, usually placed under the title, to convey the main message of the article. In our analysis, we considered just the headlines and the snippets since they are freely available to the readers and have a bigger impact than the whole article. We downloaded the headlines and the snippets of the articles using The New York Times API and then, following Bybee et al. (2020) and Kalamara et al. (2022), eliminated several sections that were not pertinent for the analysis, that is, not containing relevant information that might affect the financial markets (see Table 1). Articles published after 4:00 pm, when the stock exchanges were closed, were assigned to the next day. Also, articles published over the weekend or on days in which the New York Stock Exchange was closed were assigned to the next working day (usually the next Monday).

2.2 Topic analysis: Latent Dirichlet Allocation

To extract the topics (the subjects, the themes) of the articles, we use Latent Dirichlet Allocation (LDA), an unsupervised machine learning technique introduced by Blei et al. (2003) for text mining. The power of LDA resides in its ability to automatically identify the topics in the articles without the need for human intervention, that is, without the need to read them by an experienced reader. LDA assumes that each document, which is a newspaper article in our case (or, more precisely, the headline and the snippet of the article), is made up of various words, and that the set of all documents form what we call the corpus. In this setting, topics are latent (nonobservable) probability distributions over words, and words with the highest weights are normally used to assign meaningful names to the topics. Of course, this somehow subjective labelling of the topics does not affect in any way the analysis and is used to help in the interpretation of the results. LDA supplies the most probable topics related to each article.

Before applying LDA, our raw text data needs to be ‘cleaned’, that is, to be preprocessed. For this, we follow the same steps of Hansen et al. (2018). First of all, the preprocessing involves converting all words in the corpus into lowercase and removing any punctuation marks. Next, it requires the removal of all ‘stop’ words such as ‘a’, ‘you’, ‘themselves’, etc., which are repeated in the documents without providing relevant information on the topics. The remaining words are then stemmed to their base root. For instance, the words ‘inflationary’, ‘inflation’, ‘consolidate’, and ‘consolidating’ are converted into their stems, which are ‘inflat’ and ‘consolid’, respectively. Thus, the stems are ordered according to the *term frequency-inverse document frequency* (tf-idf) index. This index grows with the number of times a stem appears in a

document, and decreases as the number of documents containing that stem increases. It serves to eliminate common and unusual words. All stems with a value of 12,000 or lower have been disregarded. Overall, we came out with a corpus containing a total number of 29,225 articles, 502,173 stems, and 10,314 unique stems.

After preprocessing the data, we carried out the LDA analysis on the 'cleaned' corpus, fixing at 60 the total number of topics, and setting the hyperparameters of the Dirichlet priors following the suggestions of Griffiths and Steyvers (2004), as in Hansen et al. (2018). To obtain a sample from the posterior distribution, we then considered two runs of the Markov chain Monte Carlo Gibbs sampler, each one providing 1,000 draws, using a burn-in period of 1000 iterations and a thinning interval of 50.

Tables 2 and 3 show, for each of the 60 topics, the first six words with the highest (posterior) probability. That is, for each topic, word 1 is the word (stem) with the highest probability in that topic, word 2 is the word (stem) with the second highest probability in that topic, and so on. On the basis of the probability distribution of words in a topic, we are able to somehow interpret it and then assign it a tag. For instance, we assigned the tag 'coronavirus' to topic 29 since, for this topic, the words (stems) with the highest probability are 'coronaviru', which has a probability of 0.217, 'test', which has a probability of 0.057, 'pandem', which has a probability of 0.053, and 'viru', which has a probability of 0.051. In this way, we see that topics related to the economy and the financial markets are those numbered 3, 10, 36, 46, and 51. Topics related to politics are those numbered 12, 13, 15, 24, 28, 30, 31 and 35. Whereas topics related to the international economy and political conditions include those numbered 8, 14, 23, 33, 44, 48, and 53. We should remark that we carried out the LDA analysis fixing at 60 the number of topics since, with this number, we were able to clearly distinguish between the 'coronavirus' and 'trade war' topics. A larger number of topics supplies several topics related to the coronavirus pandemic (and not just one), whereas a lower number of topics, such as 40, for instance, does not clearly distinguish the 'trade war' topic from the others. In other words, we selected the number of topics providing the most understandable results, as, for instance, in Hansen et al. (2018) and Soto (2021). Alternatively, it would have been possible to select the number of topics in a more automatic way by using, for instance, the measures proposed by Hasan et al. (2021), which they called Normalized Absolute Coherence (NAC) and Normalized Absolute Perplexity (NAP).

In addition to the above probability distributions of words characterizing each topic, the LDA analysis also provides the topic distribution for each document in the corpus, that is, it supplies the most probable topics associated with each article of The New York Times. These distributions will be used to obtain the daily distributions of topics over the period under scrutiny. In particular, we will consider the daily probability of each topic, $P_{i,t}$, where subscript i refers to the topic and subscript t to the day. This text measure will be used in Sect. 2.4 to construct our topic-specific uncertainty indices.

2.3 Uncertainty analysis: Word Embedding and K-Means

In our situation, an article may convey a *certain* or an *uncertain* sentiment about a topic. This uncertain sentiment of an article will be deduced by using Word Embedding (with

the Skip-gram model) and K-Means. These algorithms will provide a list of words, having a meaning similar to that of the word 'uncertainty', which will operate as an *uncertainty dictionary*. This, in turn, will be employed to measure the uncertainty present in each article and so to build a daily uncertainty index.

Word Embedding, introduced by Mikolov et al. (2013), is a continuous vector representation of words in a suitable low-dimensional Euclidean space, which aims to capture syntactic and semantic similarities between words, associating words with a similar meaning with vectors that are closer to each other, that is, that are in the same region of the space. Usually, this can be implemented by adopting either the Common Bag Of Words (CBOW) model or the Skip-gram model. The main idea of these models is the possibility to extract a considerable amount of the meaning of a word from its *context* words, that is, from the words surrounding it. For instance, consider the following two sentences:

the economy experienced a period of increasing *uncertainty* about the growth capacity;

the economy experienced a period of increasing *fears* about the growth capacity.

Here, the words 'uncertainty' and 'fears' have a similar meaning, which is related to doubt and worry. Both words are preceded by 'the economy experienced a period of increasing' and are followed by 'about the growth capacity'. For our purposes, to carry out the Word Embedding we adopt the Skip-gram model as introduced by Mikolov et al. (2013). The basic idea of this model is to create a dense vector representation of each word that is good at predicting the words that appear in its context. This involves the use of a neural network designed to predict context words on the basis of a given *center* word.

Before proceeding with the Word Embedding, using the Skip-gram model, for the words in the articles of the relevant sections of The New York Times, we first need to preprocess the raw text data, though in a different manner than we did for LDA. Now, words are not stemmed since we could lose semantic differences between some of them. Instead, we now single out bigrams, that is, pairs of consecutive words such as, for instance, 'south_korean' or 'defense_secretary', that jointly bear a particular meaning or idea. Bigrams, that is, the two words forming it, are considered as a single token, that is, as if they were a single word. In the analysis, we considered all bigrams appearing with a frequency higher than 50. We fixed this threshold since it allows us to capture many relevant bigrams, although excluding those with relatively low frequency. A different strategy might consist in selecting the number of bigrams following some intrinsic evaluators of Word Embedding (Wang et al. 2019), such as word similarity, word analogy, concept categorization, outlier detection, and QVEC. Thus, we discarded from the analysis all articles that do not normally have an effect on financial markets, such as for instance, articles on local crime or on New York local news, which might bias the results. Specifically, we eliminated all the articles whose main topic, that is, whose highest LDA topic probability is relative to one of the following topics: 0, 5, 6, 7, 8, 9, 11, 18, 21, 22, 27, 28, 34, 35, 37, 43, 44, 48, 57 and 59. After this cleaning, we remained with a corpus of 19,713 articles and 342,038 tokens (which are either bigrams or single words). On the cleaned set of 19,713 articles, we considered Word Embedding, using the Skip-gram model, with a hidden layer of $H = 200$ elements and a context window of size 10 on each side of the center

word (we also tried a hidden layer of 100 and 150 elements, and a context window of size 5 and 8). We implemented it using Word2Vec of the Gensim Python library. This embedding has been carried out for all unique terms (words) and all identified bigrams in the selected set of articles, to obtain, for each token (word or bigram), a dense vector of dimension H.

Then, to identify tokens with a similar meaning, we performed a K-Means clustering on the dense vectors thus obtained. K-Means is an unsupervised machine learning technique that clusters similar objects, which are in some sense close to each other, in a set of disjoint clusters (MacQueen 1967). After some investigations in which we tried different combinations of the number of elements of the hidden layer, the context window size, and the number of clusters, we fixed the number of clusters at 120. The chosen combination and, in particular, the chosen number of clusters is the one that provides, with respect to the purposes of our investigation, the most meaningful results in terms of semantic similarities.

Having obtained clusters of vectors related to tokens (words or bigrams) with similar meanings, we went on (as in Soto (2021)) to identify those clusters containing words related to *uncertainty*. Precisely, we considered the clusters containing the words 'fear', 'fears', 'worries', 'uncertain', and 'uncertainty'. Tables 4, 5, 6, 7 and 8 show the words that appear in these clusters. We can note that the cluster containing the word 'uncertainty' mainly includes words related to the trade war between China and the US, whereas the cluster containing the word 'worries' mainly includes words related to stock markets. It should also be noted that a number of clusters smaller than 120 leads to clusters containing more than one of these five uncertainty words but also containing many words that are not of interest. As we said, we emphasized the interpretability of the results. Alternatively, the number of clusters might have been selected by adopting, for instance, the Elbow Criterion as in Haider et al. (2020). All the words in these five clusters were merged together to build a list of words to be used as a dictionary of words related to the sentiment of uncertainty. For our purposes, this uncertainty dictionary seems to be better than other pre-established uncertainty dictionaries, such as that of Loughran and Mcdonald (2011), since it is tailored to our particular text data. Indeed, the dictionary of Loughran and Mcdonald (2011) is principally used for very large financial data sets, whereas our text data is of a different nature covering a wider range of arguments.

With our uncertainty dictionary, we are now in a position to set up a daily uncertainty index for the US economy, which can be used to investigate the effect of uncertainty about the US economy on the financial markets. To construct this index, we first count the number of words in the uncertainty dictionary that are present in each article. The daily sum of uncertainty words, over all articles of a particular day t, is indicated by U_t. A daily uncertainty score S_t can then be obtained by dividing U_t by the total number N_t of words present in the articles that day:

$$S_t = U_t / N_t. \tag{1}$$

Our daily *US uncertainty index* is then given by

$$D_t = 100 \cdot \frac{S_t}{\frac{1}{M}\sum_{m=1}^{M} S_m}, \tag{2}$$

where M is the number of days of the period under study. Figure 1 shows the evolution of our US uncertainty index compared with the S&P 500 closing price index. The three peaks over a value of 125 of the moving average (with a 9-day rolling window) of the US uncertainty index correspond to important drops in the S&P 500 index.

2.4 Topic-specific uncertainty measures

Following Mamaysky (2023), we build topic-specific sentiment measures by multiplying the daily topic probabilities by the daily uncertainty index. In our case, the sentiment index is given by the daily US uncertainty index obtained through Word Embedding and K-Means clustering. Thus, to measure the uncertainty related to specific topics, we consider the following *topic-specific uncertainty indices*,

$$T_{i,t} = P_{i,t} \cdot D_t, \tag{3}$$

where subscript i indicates a specific topic and subscript t refers to a specific day.

Figure 2 shows the evolution of two topic-specific uncertainty indices, specifically of the 'coronavirus' and 'trade war' uncertainty indices. Similarly, Figs. 3, 4 and 5 show the evolution of the 'Brexit', 'economic-Fed', and 'climate change' uncertainty index, respectively. From these behaviors it is immediate to notice that the peaks of the 'trade war' uncertainty index during 2019 correspond to drops in the S&P 500 closing price index, whereas the huge increase of the 'coronavirus' uncertainty index in the first months of 2020 corresponds to a historic drop in the S&P 500 index.

3 Uncertainty in news and financial markets volatility

To quantify how much of the behavior of some US financial indices such as the S&P 500 index, the Dow Jones index, the Nasdaq Composite index, the VIX index, and the US 10-year Treasury bond yields, can be explained by our topic-specific uncertainty indices, we estimated various Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) models (Nelson 1991). As before, we considered the interval from 2 January 2019 to 1 May 2020, which is characterized by a period of extremely high volatility that goes from February 2020 to the end of our sample. The choice of a model of the ARCH family is suggested by the desire to explain phases of high and low volatility in the interval under study. An advantage of the EGARCH model over the more standard GARCH model is its ability to capture asymmetric behaviors, also known as leverage effects, that is, to model the asymmetric effect on the volatility of good and bad news. Specifically, a positive leverage means that high positive returns are followed by larger increases in volatility than in the case of negative returns of the same size, whereas a negative leverage means that high negative returns are followed by larger increases in volatility than in the case of positive returns.

In particular, for a given financial index f, let us consider the returns

$$\Delta C_{f,t} = \frac{C_{f,t} - C_{f,t-1}}{C_{f,t-1}} \cdot 100, \tag{4}$$

where $C_{f,t}$ is the daily closing price of the financial index f at time t. We first investigate how much of the mean and volatility of the S&P 500 returns can be explained by each of our topic-specific uncertainty indices: 'trade war', 'coronavirus', 'Brexit', 'climate change' and 'economic-Fed'. To do this, we estimated a separate EGARCH model for each of these topic-specific uncertainty indices, considering the same combination of explanatory variables used by Mamaysky (2023) in his contemporaneous regressions. Precisely, we estimated the following EGARCH(1,1) model for the S&P 500 returns $\Delta C_{S,t}$ and for each of our topic-specific uncertainty indices:

$$\begin{aligned} \Delta C_{S,t} = \; & b_0 + b_1 \Delta C_{S,t-1} + b_2 T_{i,t} + b_3 T_{i,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) \\ & + b_4 \mathrm{VIX}_{t-1} + \theta \epsilon_{t-1} + \epsilon_t, \end{aligned} \tag{5}$$

$$\begin{aligned} \ln \sigma_t^2 = \; & \omega + b_5 T_{i,t} + b_6 T_{i,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) + b_7 \mathrm{VIX}_{t-1} \\ & + \beta \ln \sigma_{t-1}^2 + \alpha \mid \frac{\epsilon_{t-1}}{\sigma_{t-1}} \mid + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}}. \end{aligned} \tag{6}$$

The mean equation in (5), measuring the influence of the explanatory variables on the mean returns of the S&P 500, includes as explanatory variables: the ith topic-specific uncertainty index $T_{i,t}$, the product of this index and the difference between the lag value VIX_{t-1} and the mean value $\overline{\mathrm{VIX}}$ of the VIX index, and the lag value of the VIX index. Similarly for the conditional variance equation with asymmetric effects, given in (6), which measures the effect of the explanatory variables on the volatility in the returns of the S&P 500. In the equations, ϵ_t refers to the zero mean and unit variance independent and identically distributed error term (ARCH error), whereas σ_t indicates the conditional variance (GARCH term). Moreover, the coefficient ω is a constant, β is the GARCH coefficient (persistence term), α is the coefficient of the ARCH term, and γ indicates the asymmetric or leverage effect.

Table 9 shows the estimates and standard errors of the parameters of the EGARCH(1,1) model in Eqs. (5) and (6), for each of the five topic-specific uncertainty indices used as an explanatory variable in the models.

The figures show the effect of a unit increase in a given topic-specific uncertainty index on the mean and volatility of the returns of the S&P 500. As expected, we see that the 'trade war' and 'coronavirus' uncertainty indices have a negative effect on the mean, and a positive effect on the volatility, of the returns of the S&P 500, though the volatility coefficient of the 'trade war' uncertainty index is not significant. Table 9 also shows that a rise in the 'Brexit' uncertainty index implies an increase in the mean of S&P 500 returns; in other words, uncertain news about Brexit did not cause negative effects on these returns. On the other hand, the 'climate change' uncertainty index seems to have a small negative effect on the mean returns of the S&P 500. Furthermore, the 'economic-Fed' uncertainty index, which accounts for news on the actions of the Fed and of the US government, seems to be positively associated

with both the mean and the volatility of the S&P 500 returns. Indeed, this uncertainty index seems to incorporate news about possible future actions of the Fed and the US government in addressing economic turmoils during periods of great uncertainty. A greater value of this index might be due to the negative economic scenarios associated with the actions of the Fed and the US government, which are, these latter, immediately absorbed by the markets with changes in companies' stock value.

As we can see from the results reported at the bottom of Table 9, the models related to the 'coronavirus', 'trade war', and 'economic-Fed' uncertainty indices passed numerous tests, including the weighted Ljung-Box test, which means that the standardized residuals are not autocorrelated, and the weighted ARCH LM test, which says that the EGARCH(1,1) models are correctly fitted. The two EGARCH(1,1) models with the best fit are those for the 'coronavirus' and 'trade war' uncertainty indices. In comparison with the other three models, these two uncertainty indices obtain the highest log-likelihood and the smallest values for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These findings seem in agreement with the graphs in Fig. 2, which suggest a negative correlation between the 'trade war' and 'coronavirus' uncertainty indices and the mean returns of the S&P 500. In particular, the 'trade war' uncertainty index seems to explain much of the behavior of the S&P 500 during 2019, whereas the 'coronavirus' uncertainty index seems to best explain the beginning of 2020. Overall, these two indices seem to do better than the other three uncertainty indices in explaining the returns of the S&P 500 from the beginning of 2019 to the end of April 2020.

To deepen the investigation on the relationship between uncertainty in the news and behavior of the financial markets, we estimated some other EGARCH models to study the joint effect of the 'coronavirus' and 'trade war' uncertainty indices on the returns of some US financial indices, in particular the S&P 500 index, the Dow Jones index, the Nasdaq Composite index, the VIX index as well as the US 10-year Treasury bonds yields. Precisely, for each of these five financial indices we considered the following EGARCH(1,1) model:

$$\begin{aligned}\Delta C_{f,t} = \; & b_0 + b_1 \Delta C_{f,t-1} + b_2 T_{\mathrm{C},t} + b_3 T_{\mathrm{W},t} + b_4 T_{C,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) \\ & + b_5 T_{W,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) + b_6 \mathrm{VIX}_{t-1} + \theta \epsilon_{t-1} + \epsilon_t,\end{aligned} \tag{7}$$

$$\begin{aligned}\ln \sigma_t^2 = \; & \omega + b_7 T_{\mathrm{C},t} + b_8 T_{\mathrm{W},t} + b_9 T_{C,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) \\ & + b_{10} T_{W,t}(\mathrm{VIX}_{t-1} - \overline{\mathrm{VIX}}) + b_{11} \mathrm{VIX}_{t-1} \\ & + \beta \ln \sigma_{t-1}^2 + \alpha \left| \frac{\epsilon_{t-1}}{\sigma_{t-1}} \right| + \gamma \frac{\epsilon_{t-1}}{\sigma_{t-1}},\end{aligned} \tag{8}$$

where $T_{\mathrm{C},t}$ and $T_{\mathrm{W},t}$ refer to the 'coronavirus' and 'trade war' uncertainty indices, respectively, and $\Delta C_{f,t}$ indicates the returns of the financial index f at time t.

Table 10 shows the estimates and standard errors of the parameters of the EGARCH(1,1) model in Eqs. (7) and (8), for each of the five financial indices used for the dependent variable in the mean equation.

As expected, we see that both the 'coronavirus' and 'trade war' uncertainty indices have a negative effect on the mean and a positive effect on the volatility of the returns

of the S&P 500. In particular, we notice that an increase in the 'trade war' uncertainty index has a greater negative effect on the mean returns of the S&P 500 than an increase in the 'coronavirus' uncertainty index. Let us also observe that the 'coronavirus' uncertainty index has a negative effect on the mean returns of the Nasdaq, but not on that of the Dow Jones, and vice-versa for the 'trade war' uncertainty index. Moreover, we see that the mean returns of the VIX are positively affected by the 'coronavirus' and 'trade war' uncertainty indices. Lastly, as far as the 10-year US Treasury bond yields are concerned, the results show that an increase in the 'coronavirus' and 'trade war' uncertainty indices leads to a decrease in their mean returns. In line with common opinion, we can reasonably argue that investors may see US bonds as a safe refuge during periods of high uncertainty.

The bottom of Table 10 shows that the models for the S&P 500, the VIX, and the 10-year US Treasury bond yields passed both the weighted Ljung-Box test, which indicates that the standardized residuals are not autocorrelated, and the weighted ARCH LM test, which means that the EGARCH process is correctly fitted. By far, the EGARCH(1,1) model with the best fit is that for the S&P 500. Comparing it with the other four models, this model has the highest log-likelihood and the smallest values for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

4 Conclusions

In this paper we use unsupervised machine learning techniques to construct text measures able to explain recent past movements in US financial markets. Our raw text data are the headlines and snippets of the articles of The New York Times from 2 January 2019 to 1 May 2020. We first use LDA to infer the content (topics) of the articles and thus to obtain daily indices on the presence of these topics in The New York Times. Then we use Word Embedding (implemented with the Skip-gram model) and K-Means to construct a daily uncertainty measure. Thus, we combine all these measures to obtain daily topic-specific uncertainty indices. In particular, we obtain five uncertainty indices related to news about 'coronavirus', 'trade war', 'Brexit', 'economic-Fed' and 'climate change', capturing the daily degree of uncertainty in these topics.

To quantify how much of the behavior of the S&P 500 index can be explained by uncertainty in the news, we estimated an EGARCH(1,1) model for each of our five topic-specific uncertainty indices. We verify that the 'coronavirus' and 'trade war' uncertainty indices are negatively associated with the mean and positively associated with the volatility of the returns of the S&P 500. Also, we find that the 'climate change' and 'economic-Fed' uncertainty indices are negatively and positively, respectively, associated with the mean of the S&P 500 returns. This suggests that news about economic measures of the Fed and the US government has a positive effect on the S&P 500 in days of uncertainty. Overall, we can argue that the 'trade war' uncertainty index explains much of the behavior of the S&P 500 returns during 2019, whereas the 'coronavirus' uncertainty index explains most of the movements of the S&P 500 index during the first four months of 2020.

To further investigate how much these two uncertainty indices explain the behavior of the US financial markets, we estimated, using these two indices as explanatory

variables, some other EGARCH(1,1) models, one for each of the following financial indices (as a dependent variable): the S&P 500, the Nasdaq, the Dow Jones, the VIX and the US 10-year Treasury bond yields. We find that the 'coronavirus' and 'trade war' uncertainty indices have a negative effect on the mean and a positive effect on the volatility of the returns of the S&P 500. We also find that these two uncertainty indices have a positive effect both on the mean and the volatility of the returns of the VIX index.

Future research might address some issues raised by the use of the headlines and the snippets instead of the (lacking) full text of the articles in The New York Times. A better uncertainty dictionary could reasonably be obtained by considering a larger set of articles, maybe considering more than one newspaper. Though the analysis and the model were quite successful in explaining the gathered data and some of the reactions to the first wave of the COVID-19 pandemic, we highlight that the data covers a limited and peculiar time period, daily from January 2019 to May 2020. In the future, it would be interesting to repeat the analysis over a longer time period. Definitely, it must be underlined that in this paper we carried out an ex-post analysis. It might be interesting to investigate the ability of our method to identify meaningful topics in real-time, in particular, a coronavirus topic at the beginning of the pandemic in February and March 2020. This is an important point since it would allow the creation of real-time indicators that might be used in forecasting tasks. From a methodological point of view, it should also be explored the use of other machine learning methods for the construction of text measures such as Dynamic Topic Models (Blei and Lafferty 2006) and Support Vector Machines. Also, it might be interesting to explore other sentiments other than uncertainty. For instance, future investigations might consider a positive–negative sentiment index using the Loughran-McDonald dictionary (Loughran and Mcdonald 2011), or using a positive–negative dictionary based on our corpus, following the methodology of Soto (2021) used in this paper. Moreover, it might also be interesting to investigate the FinBert procedure proposed by Huang et al. (2023), which classifies phrases as positive or negative and seems to outperform, at least in some contexts, the Loughran-McDonald dictionary and other machine learning methods. Finally, more sophisticated GARCH-MIDAS models could be used to incorporate, as explanatory variables, macroeconomic and other variables sampled at different frequencies, as well as regime-switching models could be used in the study of the impact of news on financial markets.

Appendix

Table 1 List of sections of The New York Times not considered in the analysis

arts and leisure, at home, book, briefing, corrections, crosswords and games, culture, dining, express, fashion, fashion and style, food, games, gender, graphics, health, insider, learning, letters, live, magazine, metropolitan, movies, multimedia / photos, New York, none, obit, obituaries, parenting, photo, reader center, smarter living, real state, society, special section, sports, style, styles, Sunday review, t magazine, t magazine / art, t magazine / fashion and beauty, tstyle, the learning network, the weekly, theater, times insider, travel, weekend and well.

Table 2 Topic descriptions for the LDA analysis. The table shows the first six words with the highest (posterior) probability for each of the first thirty topics

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
0. Sexual crime	claim	accus	abus	sexual	file	assault
	0.057	0.047	0.034	0.031	0.022	0.021
1. Face / threat	face	critic	threat	challeng	remain	potenti
	0.118	0.045	0.044	0.04	0.04	0.027
2. Need/ help	need	know	will	help	car	want
	0.13	0.086	0.049	0.048	0.03	0.028
3. Economy / Fed	economi	econom	bank	cut	rate	feder
	0.068	0.062	0.05	0.043	0.037	0.029
4. Executive chief	chief	execut	mr	former	role	head
	0.056	0.047	0.047	0.037	0.031	0.029
5. Black culture	black	histori	cultur	celebr	look	photo
	0.053	0.041	0.024	0.023	0.019	0.018
6. Effort / move	tri	move	effort	part	canada	stop
	0.07	0.066	0.052	0.045	0.028	0.027
7. Crime investigation	charg	case	prison	former	prosecutor	crime
	0.059	0.048	0.035	0.033	0.032	0.028
8. Politics Spain Sudan	power	leader	call	polit	anti	countri
	0.081	0.067	0.053	0.048	0.035	0.033
9. Time	year	last	month	decad	nearli	ago
	0.227	0.087	0.073	0.045	0.029	0.029
10. Labour	work	govern	worker	job	pay	employe
	0.111	0.071	0.063	0.056	0.034	0.028
11. Immigration	border	immigr	migrant	wall	mexico	famili
	0.084	0.057	0.045	0.038	0.037	0.027
12. Democratic party	democrat	biden	debat	sander	candid	berni
	0.112	0.086	0.086	0.06	0.042	0.037
13. White house	hous	trump	white	presid	democrat	aid
	0.184	0.122	0.096	0.059	0.033	0.029
14. Iran	iran	storm	flood	iranian	hit	strike
	0.083	0.026	0.021	0.021	0.02	0.017
15. Trump Ukraine	presid	trump	ukrain	lawyer	impeach	mr
	0.07	0.07	0.049	0.033	0.032	0.031
16. Election campaign	question	campaign	democrat	ask	candid	iowa
	0.075	0.062	0.051	0.045	0.044	0.04
17. Airplane crash	india	crash	air	boe	travel	plane
	0.042	0.033	0.026	0.026	0.026	0.022

Table 2 (continued)

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
18. Education	school	student	colleg	children	public	parent
	0.062	0.056	0.045	0.035	0.031	0.028
19. Research	found	find	research	human	scientist	studi
	0.047	0.042	0.039	0.036	0.034	0.033
20. Tech companies	compani	use	tech	data	big	giant
	0.068	0.057	0.056	0.037	0.036	0.029
21. Multimedia	show	video	play	watch	servic	game
	0.102	0.038	0.029	0.027	0.026	0.026
22. Justice	court	rule	case	suprem	judg	justic
	0.106	0.082	0.038	0.036	0.036	0.034
23. North Korea	north	meet	south	talk	korea	end
	0.064	0.059	0.057	0.048	0.044	0.04
24. Donald Trump	trump	presid	administr	donald	alli	tweet
	0.449	0.309	0.045	0.02	0.01	0.009
25. Future	will	week	next	come	set	expect
	0.262	0.077	0.064	0.063	0.031	0.031
26. Law	law	bill	control	gun	limit	congress
	0.049	0.048	0.047	0.042	0.037	0.033
27. Gender	women	famili	woman	men	die	life
	0.088	0.047	0.037	0.035	0.031	0.027
28. Politics	plan	warren	elizabeth	propos	seek	offer
	0.138	0.062	0.041	0.038	0.025	0.023
29. Coronavirus	coronaviru	test	pandem	viru	spread	outbreak
	0.217	0.057	0.053	0.051	0.037	0.037

Table 3 Topic descriptions for the LDA analysis. The table shows the first six words with the highest (posterior) probability for each of the last thirty topics

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
30. Election	elect	vote	voter	win	result	parti
	0.142	0.067	0.049	0.043	0.037	0.028
31. Politics	polit	turn	fight	governor	line	point
	0.118	0.055	0.054	0.041	0.038	0.031
32. Money	million	billion	money	fund	busi	rais
	0.053	0.052	0.049	0.049	0.045	0.039

Table 3 (continued)

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
33. Brexit	minist	prime	brexit	may	britain	european
	0.077	0.065	0.051	0.05	0.042	0.039
34. Attack / Shooting	kill	attack	shoot	peopl	polic	taliban
	0.086	0.077	0.038	0.036	0.033	0.024
35. Political groups	right	group	far	parti	left	support
	0.107	0.053	0.053	0.052	0.045	0.041
36. Tax	tax	break	israel	return	give	west
	0.061	0.04	0.039	0.039	0.024	0.024
37. Health care	health	care	crisi	public	system	emerg
	0.1	0.072	0.061	0.051	0.04	0.04
38. Foreign security	offici	secur	nation	top	foreign	secretari
	0.111	0.069	0.054	0.047	0.041	0.035
39. Social news	social	news	media	facebook	ad	onlin
	0.06	0.058	0.047	0.041	0.04	0.035
40. Russian investigation	report	gener	investig	russia	mueller	russian
	0.088	0.072	0.06	0.055	0.043	0.037
41. Death toll	death	record	number	rise	show	tip
	0.081	0.054	0.036	0.031	0.024	0.019
42. American nation	state	unit	american	nation	offici	address
	0.279	0.108	0.097	0.032	0.028	0.028
43. Story / book	stori	love	read	tell	week	book
	0.052	0.036	0.034	0.034	0.032	0.025
44. France space	franc	land	space	french	trip	light
	0.028	0.025	0.024	0.021	0.016	0.016
45. World	world	countri	around	across	america	fear
	0.124	0.114	0.044	0.043	0.037	0.028
46. Stock market	market	stock	compani	price	oil	fall
	0.058	0.037	0.035	0.034	0.033	0.029
47. Verbs	want	look	listen	daili	live	let
	0.086	0.043	0.034	0.028	0.025	0.023
48. Syria	forc	american	militari	war	syria	turkey
	0.061	0.06	0.06	0.04	0.033	0.022
49. Medicine	drug	use	doctor	patient	peopl	hospit
	0.044	0.04	0.037	0.033	0.032	0.031
50. Home	home	citi	stay	peopl	commun	resid

Table 3 (continued)

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
	0.087	0.083	0.038	0.035	0.031	0.03
51. Trade war	china	trade	deal	war	Chines	Talk
	0.17	0.085	0.066	0.058	0.052	0.034
52. Impeachement	senat	impeach	republican	democrat	trial	trump
	0.122	0.101	0.094	0.067	0.047	0.03
53. Hong kong protest	protest	hong	kong	polic	govern	thousand
	0.11	0.06	0.06	0.042	0.029	0.026
54. Climate change	chang	climat	fire	california	australia	water
	0.135	0.08	0.076	0.054	0.031	0.017
55. Verbs / adjectives	much	cannot	may	good	problem	better
	0.047	0.044	0.044	0.044	0.042	0.034
56. Day	day	quotat	brief	friday	wednesday	thursday
	0.24	0.101	0.07	0.042	0.041	0.038
57. Food	close	food	open	busi	bring	industri
	0.048	0.037	0.031	0.025	0.02	0.018
68. Verbs	can	help	keep	save	thing	learn
	0.165	0.089	0.061	0.034	0.029	0.027
59. New York	time	york	report	follow	cover	journalist
	0.205	0.078	0.04	0.026	0.026	0.021

Table 4 List of words in the cluster containing the word 'fear'

anxious, anywhere, battling, belt, born, brutal, civilians, communist, contagion, crisis, deep, fake_news, fear, feels, fighting, fingers, girl, greatest, indians, isis, isolation, italy, landslide, latin_america, lockdown, locked, looks_like, memories, neighbors, nightmare, outrage, poland, react, relative, revolution, shame, siege, solidarity, suffers, test, thailand, tour, tradition, trauma, turns, upheaval, war_ii, west, widening.

Table 5 List of words in the cluster containing the word 'fears'

analysts, bond_yields, central_banks, climb, damage, drop, exports, factories, fears, fell, financial_markets, fueled, gas, grew, growing, higher, highest, increase, increasing, oil, oil_prices, plunge, policymakers, prices, producers, rate, rattled, rise, rising, slide, slowdown, slowing, slows, slump, spike, supply, tourism, tumbled, worsening.

Table 6 List of words in the cluster containing the word 'worries'

central_bank, cut_interest, cut_rates, economic, economy, fed, federal_reserve, global, growth, interest_rates, investors, markets, rates, recession, stocks, worries.

Table 7 List of words in the cluster containing the word 'uncertain'

accord, agreed, alternative, approaching, backs, backstop, bloc, blow, boris, brinkmanship, brussels, closer, collision_course, complicate, compromise, corbyn, customs, deadline, deepening, europeans, extending, failed, failure, fate, forge, gives, grant, guarantee, heads, jan, john_bercow, last_ditch, likely, limbo, looming, macedonia, maneuver, mideast, nears, negotiating, obstacles, oct, paris_climate, persuade, pound, promises, prospect, quick, rather, rebels, remain, reverse, shinzo_abe, stalemate, stamp, step, suspend_parliament, suspension, throws, tries, two_sides, uncertain, unpredictable, vacuum, vowed, wall, yearlong.

Table 8 List of words in the cluster containing the word 'uncertainty'

chinese_goods, goods, mexico, negotiations, negotiators, progress, tariff, tariffs, trade, trade_deal, trade_talks, trade_war, uncertainty.

Table 9 Estimates and standard errors (in parentheses) of the parameters of the EGARCH(1,1) model in Eqs. (5) and (6), for each of the five topic-specific uncertainty indices

	Trade War	Coronavirus	Brexit	Climate	Economic-Fed
b_0	-0.10^{**}	-0.16^{***}	0.61^{***}	0.01^{***}	-0.04^{***}
	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)
b_1	0.83^{***}	0.17^{***}	-0.53^{***}	-0.53^{***}	-0.53^{***}
	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)
b_2	-0.10^{***}	-0.01^{***}	0.12^{***}	-0.02^{***}	0.28^{***}
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
b_3	-0.01	-0.00^{***}	0.02^{***}	-0.00^{***}	-0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
b_4	0.02^{***}	0.01^{***}	-0.04^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
θ	-0.87^{***}	-0.20^{***}	0.18^{***}	0.18^{***}	0.17^{***}
	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
ω	-0.56	-0.45^{***}	3.29^{***}	3.28^{***}	3.28^{***}
	(0.34)	(0.00)	(0.00)	(0.00)	(0.00)
b_5	0.09	0.06^{***}	0.08^{***}	-0.05^{***}	0.16^{***}
	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)
b_6	0.00	-0.00^{***}	-0.00^{***}	0.04^{***}	-0.05^{***}

Table 9 (continued)

	Trade War	Coronavirus	Brexit	Climate	Economic-Fed
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
b_7	0.02	0.02***	−0.21***	−0.14***	−0.21***
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
β	0.76***	0.83***	0.90***	0.90***	0.90***
	(0.10)	(0.00)	(0.00)	(0.00)	(0.00)
α	−0.34***	−0.48***	0.07***	0.04***	0.04***
	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)
γ	0.15	−0.44***	0.13***	0.11***	0.11***
	(0.11)	(0.00)	(0.00)	(0.00)	(0.00)
Log likelihood	−425.55	−416.87	−1766.02	−2698.76	−2646.47
AIC	2.61	2.56	10.59	16.14	15.83
BIC	2.76	2.71	10.74	16.29	15.98
Ljung-Box Test (p-value in parentheses)					
Lag[1]	0.01027	0.3799	$1.872e-04$	0.05225	0.001937
	(0.9193)	(0.5377)	($9.891e-01$)	($8.192e-01$)	(0.9649)
Lag[2*(p+q)+(p+q)-1][5]	0.57671	1.0545	$1.877e+00$	19.62855	1.655162
	(1.0000)	(1.0000)	($9.768e-01$)	($0.000e+00$)	(0.9937)
Lag[4*(p+q)+(p+q)-1][9]	3.20533	4.5854	$2.417e+01$	30.59862	2.760406
	(0.8571)	(0.5507)	($2.907e-10$)	($3.897e-14$)	(0.9236)
ARCH LM Test (p-value in parentheses)					
ARCH Lag[3]	0.4612	0.1345	0.01335	15.11	0.003194
	(0.4971)	(0.71379)	(0.908002)	($1.017e-04$)	(0.9549)
ARCH Lag[5]	0.5281	1.3910	0.57237	15.44	0.018647
	(0.8753)	(0.62143)	(0.862059)	($2.970e-04$)	(0.9999)
ARCH Lag[7]	0.7583	8.0744	15.10207	19.69	0.031256
	(0.9494)	(0.05051)	(0.001089)	($7.647e-05$)	(1.0000)

Each column header indicates the topic-specific uncertainty index used as an explanatory variable in the model. The dependent variable in all five models is the returns of the S&P 500

p-value: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 10 Estimates and standard errors (in parenthesis) of the parameters of the EGARCH(1,1) model in Eqs. (7) and (8), for each of the five financial indices

	S&P 500	Nasdaq	Dow Jones	VIX	Treasury yields
b_0	−0.87***	0.86***	−1.16***	1.45***	−2.19***
	(0.00)	(0.02)	(0.00)	(0.33)	(0.22)
b_1	−0.76***	0.06***	−0.60***	−0.48***	−0.87***
	(0.00)	(0.01)	(0.00)	(0.12)	(0.05)
b_2	−0.03***	−0.58***	0.08***	0.63***	−0.41**
	(0.00)	(0.00)	(0.00)	(0.08)	(0.15)
b_3	−0.17***	0.03***	−0.22***	0.45***	−0.38***
	(0.00)	(0.00)	(0.00)	(0.11)	(0.10)
b_4	−0.00***	−0.03***	−0.02***	0.02***	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
b_5	−0.02***	0.01***	−0.05***	−0.15***	−0.06**
	(0.00)	(0.00)	(0.00)	(0.03)	(0.02)
b_6	0.07***	−0.03***	0.07***	−0.22***	0.16***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
θ	0.76***	−0.38***	0.29***	0.34**	0.83***
	(0.00)	(0.00)	(0.00)	(0.12)	(0.07)
ω	0.22***	−1.88***	−1.86***	0.48***	0.21
	(0.00)	(0.01)	(0.00)	(0.04)	(0.40)
b_7	0.07***	−0.78***	−0.88***	0.09***	0.14**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.05)
b_8	0.08***	0.08***	0.15***	0.05***	0.09
	(0.00)	(0.00)	(0.00)	(0.01)	(0.06)
b_9	−0.00***	−0.34***	−0.35***	−0.00	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
b_{10}	0.01***	0.01***	0.03***	−0.01***	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
b_{11}	−0.03***	0.10***	0.10***	−0.01***	0.03
	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)
β	0.89***	1.00***	0.93***	0.87***	0.41*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.19)
α	−0.35***	−0.23***	0.22***	0.38***	−0.14
	(0.00)	(0.00)	(0.00)	(0.04)	(0.09)
γ	−0.31***	0.58***	0.39***	−0.13**	0.70***
	(0.00)	(0.00)	(0.00)	(0.05)	(0.16)
Log likelihood	−409.42	−1838.45	−1774.76	−1133.32	−811.42
AIC	2.54	11.04	10.67	6.85	4.93
BIC	2.73	11.24	10.86	7.04	5.12

Table 10 (continued)

	S&P 500	Nasdaq	Dow Jones	VIX	Treasury yields
Ljung-Box Test (*p*-value in parentheses)					
Lag[1]	1.058	0.4831	0.1755	0.6897	0.3377
	(0.3037)	(0.487)	(0.6752)	(0.4063)	(0.5611)
Lag[2*(p+q)+(p+q)-1][5]	1.640	273.7142	160.8419	1.2335	0.7844
	(0.9943)	(0.000)	(0.0000)	(0.9998)	(1.0000)
Lag[4*(p+q)+(p+q)-1][9]	5.917	447.9748	222.9947	3.3749	3.7797
	(0.2703)	(0.000)	(0.0000)	(0.8260)	(0.7415)
ARCH LM Test (*p*-value in parentheses)					
ARCH Lag[3]	0.218	0.2553	0.07989	0.1928	0.8753
	(0.6406)	(0.6134)	$(7.774e-01)$	(0.6606)	(0.3495)
ARCH Lag[5]	1.181	141.7525	36.51865	2.3831	3.3546
	(0.6802)	(0.0000)	$(1.365e-09)$	(0.3926)	(0.2423)
ARCH Lag[7]	4.256	216.2402	44.58176	3.6654	6.6470
	(0.3111)	(0.0000)	$(2.090e-11)$	(0.3974)	(0.1033)

Each column header indicates the financial index used for the dependent variable in the mean equation; the dependent variable is the returns of the index. In all five models, the explanatory variables are the 'coronavirus' and 'trade war' topic-specific uncertainty indices

p-value: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

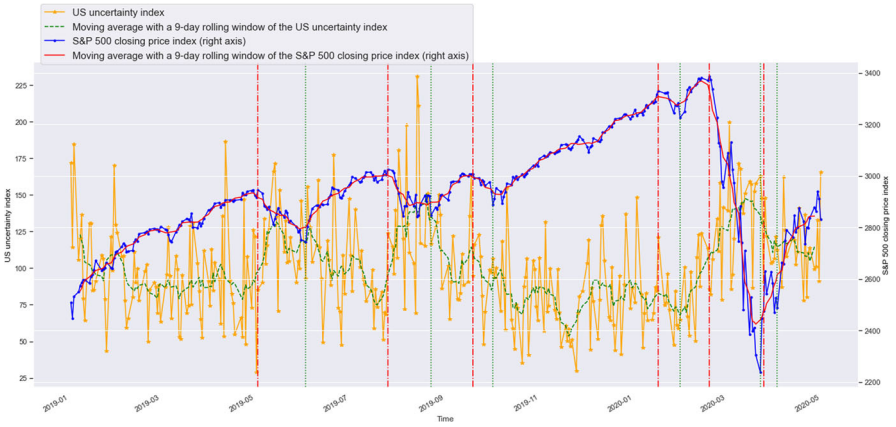

Fig. 1 Temporal evolution of the US uncertainty index. The yellow line shows the US uncertainty index obtained with the Skip-gram model. The green line represents the moving average of this index using a 9-day rolling window. The blue line shows the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines indicate some of the local minimums of the S&P 500 closing price index

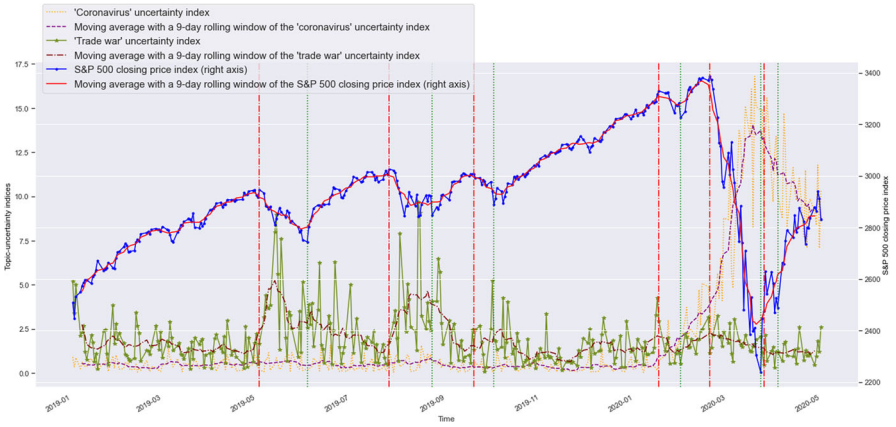

Fig. 2 Temporal evolution of the 'coronavirus' and 'trade war' uncertainty indices. The yellow line represents the 'coronavirus' uncertainty index; the purple line is the moving average with a 9-day rolling window. The green line represents the 'trade war' uncertainty index; the brown line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index

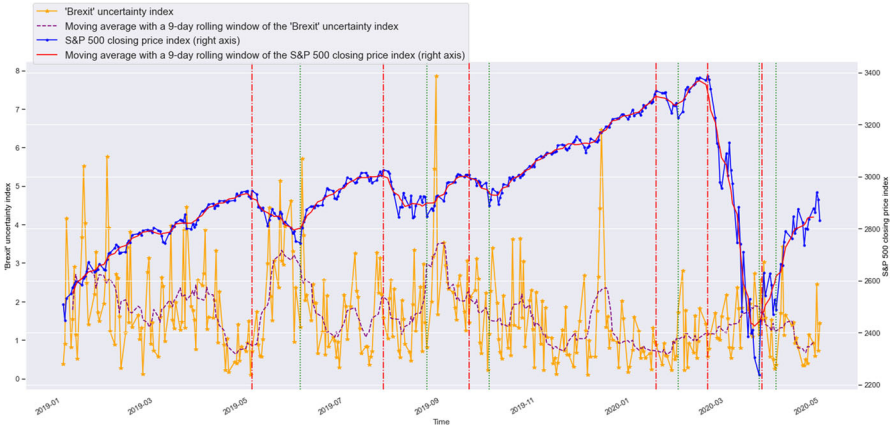

Fig. 3 Temporal evolution of the 'Brexit' uncertainty index. The yellow line represents the 'Brexit' uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index

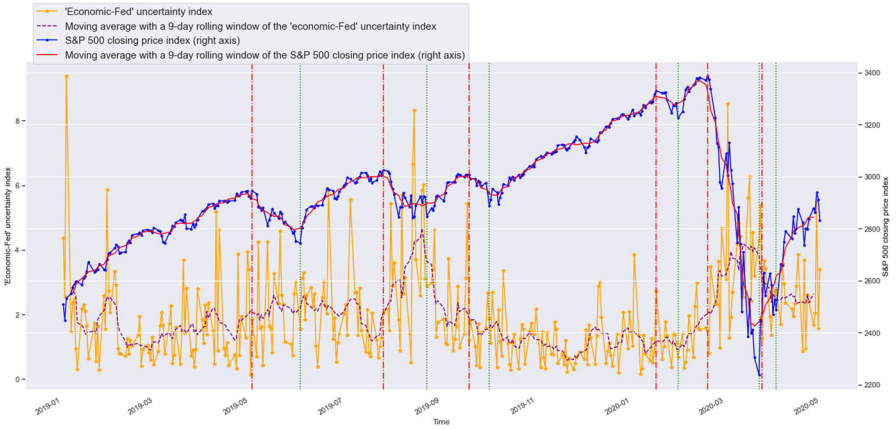

Fig. 4 Temporal evolution of the 'economic-Fed' uncertainty index. The yellow line represents the 'economic-Fed' uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing price index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index

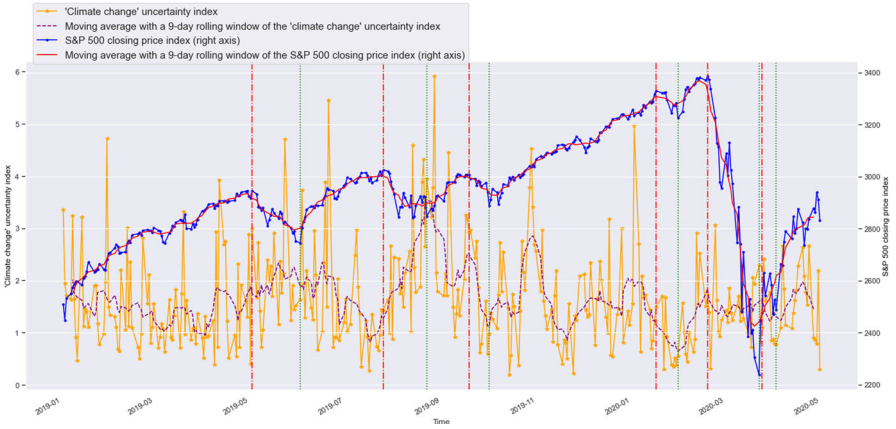

Fig. 5 Temporal evolution of the 'climate change' uncertainty index. The yellow line represents the 'climate change' uncertainty index; the purple line is the moving average with a 9-day rolling window. The blue line represents the S&P 500 closing price index; the red line is the moving average with a 9-day rolling window. The vertical dash-dotted red lines indicate some of the local maxima of the S&P 500 closing index, whereas the vertical dotted green lines represent some of the local minima of the S&P 500 closing price index

Funding Open access funding provided by Università degli Studi di Verona within the CRUI-CARE Agreement.

References

Albulescu C (2020) Coronavirus and financial volatility: 40 days of fasting and fear. arXiv preprint. arXiv:2003.04005

Ardizzi G, Emiliozzi S, Marcucci J et al (2019) News and consumer card payments. Working Paper, Banca d'Italia (1233). https://www.bancaditalia.it/pubblicazioni/temi-discussione/2019/2019-1233/index.html?com.dotmarketing.htmlpage.language=1

Azqueta-Gavaldón A, Hirschbühl D, Onorante L et al (2023) Sources of economic policy uncertainty in the euro area. Europ Econom Rev 152(104):373. https://doi.org/10.1016/j.euroecorev.2023.104373

Baker SR, Bloom N, Davis SJ et al (2020) COVID-induced economic uncertainty. National Bureau of Economic Research (w26983). https://www.nber.org/papers/w26983

Blei DM, Lafferty JD (2006) Dynamic topic models. ICML '06: proceedings of the 23rd international conference on machine learning pp 113–120. https://doi.org/10.1145/1143844.1143859

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J Mach Learn Res 3:993–1022. https://doi.org/10.5555/944919.944937

Burggraf T, Fendel R, Huynh TLD (2020) Political news and stock prices: evidence from Trump's trade war. Appl Econom Lett 27(18):1485–1488. https://doi.org/10.1080/13504851.2019.1690626

Bybee L, Kelly BT, Manela A et al (2020) The structure of economic news. Working Paper National Bureau of Economic Research (26648). https://doi.org/10.3386/w26648

Edison H, Carcel H (2021) Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. Appl Econom Lett 28(1):38–42. https://doi.org/10.1080/13504851.2020.1730748

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235. https://doi.org/10.1073/pnas.0307752101

Haider MM, Hossin MA, Mahi HR et al (2020) Automatic text summarization using Gensim Word2Vec and K-Means clustering algorithm. 2020 IEEE Region 10 Symposium (TENSYMP), pp 283–286. https://doi.org/10.1109/tensymp50017.2020.9230670

Hansen S, McMahon M (2016) Shocking language: understanding the macroeconomic effects of central bank communication. J Int Econom 99:S114–S133. https://doi.org/10.1016/j.jinteco.2015.12.008

Hansen S, McMahon M, Prat A (2018) Transparency and deliberation within the FOMC: a computational linguistics approach. Quart J Econom 133(2):801–870. https://doi.org/10.1093/qje/qjx045

Haroon O, Rizvi SAR (2020) COVID-19: media coverage and financial markets behavior-a sectoral inquiry. J Behav Experim Finance 27(100):343. https://doi.org/10.1016/j.jbef.2020.100343

Hasan M, Rahman A, Karim MR et al (2021) Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering Advances in Intelligent Systems and Computing, vol 1309 Springer, Singapore, pp 341–354. https://doi.org/10.1007/978-981-33-4673-4_27

Huang AH, Wang H, Yang Y (2023) FinBERT: a large language model for extracting information from financial text. Contem Account Res 40(2):806–841. https://doi.org/10.1111/1911-3846.12832

Jegadeesh N, Wu D (2017) Deciphering Fedspeak: The information content of FOMC meetings. https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=21466 &tk=niAkBk3N

Kalamara E, Turrell A, Redl C et al (2022) Making text count: economic forecasting using newspaper text. J Appl Econom (865). https://doi.org/10.2139/ssrn.3610770

Loughran T, Mcdonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66(1):35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

MacQueen J (1967) Some methods for classification and analysis of multivariate observations 1(14):281–297. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Mamaysky H (2023) News and markets in the time of COVID-19. Available at SSRN 3565597. https://doi.org/10.2139/ssrn.3565597
Mikolov T, Chen K, Corrado G et al (2013) Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781
Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. Econom J Econom Soc 14:347–370. https://doi.org/10.2307/2938260
Soto PE (2021) Breaking the word bank: measurement and effects of bank level uncertainty. J Financ Serv Res 59(1):1–45. https://doi.org/10.1007/s10693-020-00338-5
Wang B, Wang A, Chen F et al (2019) Evaluating Word Embedding models: methods and experimental results. APSIPA Trans Signal Inform Process 8:e19. https://doi.org/10.1017/ATSIP.2019.12
Zhu S, Liu Q, Wang Y et al (2019) Which fear index matters for predicting US stock market volatilities: text-counts or option based measurement? Physica A Statist Mechan Appl 536(122):567. https://doi.org/10.1016/j.physa.2019.122567