Imperial College London

MSc Financial Technology

Imperial College London

Business School

NEWS AND ECONOMY: A LONGITUDINAL ANALYSIS OF THE IMPACT OF FINANCIAL NEWS

Author:
01739849

Supervisor:
Prof. Stephen Hansen

August 24, 2020

Contents

Abstract

The purpose of this study is to analyze how text-based data can be use to extract useful information in order to assess global and domestic macro events. Using Natural Language Processing techniques from existing literature, I construct monthly and daily indices in order to empirically investigate the overall trend over time of Economic Uncertainty in two European countries, Italy and the United Kingdom. Moreover, this research focuses on the comparison of the behavior and approaches taken by the two countries to cope with the *Covid-19* pandemic. In order to achieve this I firstly construct an Economic Uncertainty and an Economic Policy Uncertainty index for both governments to outlines the periods with the peaks of uncertainty conveyed by the news articles. Second, I conduct a topic modeling analysis in order to extract and study how newspapers in the two countries have treated topics over time in the first half of 2020. The results show that the 2-week gap regarding the spreading of the virus, is also reflected it the change in time of the levels economic uncertainty. The topic analysis highlights the different approaches taken by the two nations. The system I propose in this paper is not pandemic-specific and can be adjusted to other countries to analyze different events.

1 Introduction

Policy Uncertainty (PU) is the risk associated to the unpredictable behavior of governments and regulators towards matters concerning different aspects ranging from monetary to fiscal policies. Over the years, this variable risk played a crucial role subsequent macroeconomic events that affected the financial system on a global scale. Recent events, such as the 2008 financial crisis or the 2014 debt crisis in Europe, have increased awareness regarding PU for researchers, on one hand, and institutions, on the other. This led to many believing that it worked as a catalyst stimulating economic recessions, with the latest trigger being the 2019 global pandemic, Balta, Fernandez and Ruscher (2013). The outbreak of the Covid-19, a novel infectious disease, generated a worldwide emergency that escalated at an exponential rate, leaving the world in a continuous state of uncertainty. Ever since November 2019, the powerful and massive transmission of news shaped the way the pandemic was perceived as it contributed to spread the risks and the uncertainties associated with it. In parallel to the outgrowing numbers, the global economy was put in question as the whole system was affected and in some cases, disrupted. In addition to disseminating news, mass media was an influential factor in the interpretation of the economic situation which in turn impacts the consumer's confidence (De Boef and Kellstedt, 2004). The unstable evaluation between positive and negative led to the rapid spread of fear and uncertainty amongst society, particularly in different European regions.

This paper has the objective to analyze the role of Policy Uncertainty and the impact of economic news on the spread of fear and risk amongst people and the evolution of uncertainty in association to its different entities, in times of a pandemic. To tackle the cited matters I construct different Economic Uncertainty indices based on the works of Baker, Bloom and Davis. In their work, they proposed a newspaper-based measure of policy uncertainty. Subsequent to their paper, I first build an Economic Uncertainty (EU) index and then the Economic Policy Uncertainty (EPU) relative to the same time period. This will effectively point out which unpredictable macroeconomic events have been, totally or in part, caused by policy uncertainty. Practically, this occurs when spikes in both indices are highly correlated. This process of constructing indices will cover two European countries, Italy and the United Kingdom. It will therefore enable a wider understanding of the macroeconomic events on both the regional and European levels. Policy Uncertainty is not an easy metric to compute, in order to test and evaluate its performance, the index has been compared to finance based measures thus efficiently proving its function as a filter for PU. Caggiano, Castelnuovo and Groshenny (2014) use stock market volatility as a proxy for PU

during their research to assess the impact of uncertainty on the economy of the United States.

The study covers a time period from January 2007 to July 2020, where I will identify and analyze the major events behind indices spikes by applying Natural Language Processing algorithm to raw text articles related to these periods. For the time frame concerning the *Covid-19* pandemic I will conduct a deeper analysis attempting the extraction of the different ways the Italian and British government have dealt with this unforeseen situation. In particular, by using a Latent Dirichlet Allocation model Blei, Ng and Jordan (2003), I will extract the time series evolution of the most discussed topics in both Italy and the UK. This will help showcase how these countries coped with this health emergency and where their attention was mainly focused. Topic Modelling was introduced by Blei et al. with the application of the Latent Dirichlet Allocation model in various sectors, making it one of the most influential statistical model for estimating low dimensional structure in discrete data. In other words, topic modeling is about extracting the thematic structure in large collections of text, annotating the documents according to that structure and finally using those annotation to visualize the documents. To be more specific, I will be going to analyze and visualize how the topics discussed varied over time as the pandemic evolved.

This paper is structured as follows: in the first section, I will be introducing the data description, methodology, models and frameworks used in the various cited papers. Next, I will be constructing and analyzing the indices for the 2007-2020 time period. The analysis is made up of a general overview of macro events corresponding to the spikes in the indices as well as the distinction between local and global crises. The third section revolves around the specifics of the *Covid-19* pandemic where I will be extracting the main differences and similarities between the two chosen countries mentioned above. This focuses on the daily EPU indices and the analysis on raw text related to this period with the objective of extracting the topics and trends that are associated with policy uncertainty. Finally, the conclusion will summarize the results obtained in the paper along with inputs for future works.

The research relies on two major pieces of literature. The first report is the work by Baker et al. (2016) where they propose a way to measure policy uncertainty from news articles. They demonstrate that the frequency of news articles related to uncertainty and policy with respect to the total number of articles published is a valid proxy for estimating PU. The second piece is 'Latent Dirichlet Allocation' by Blei et al. (2003) in which they explain the framework and technical details of their generative probabilistic model for topic extraction. This framework has been previously implemented in the economic sector by Hansen, McMahon and Prat (2018) to analyze text related to economy and policy.

2 Data and Frameworks

Before diving into the technical details of the research the jargon and terminology used throughout the paper are presented. An *article* refers to a sequence of words and punctuation whose content focus on an event discussed in the newspaper. Each article is characterized by the date of publication and its relative title. A *term* is a single word in an article. The *term-document* matrix is a matrix representing the unique terms extracted from all documents. The rows of the matrix represent a term, whereas the columns corresponds to a document. The values that populate the term-document matrix indicate the frequency of each term in each document. This approach of text mining and information retrieval does not take into consideration the semantic meaning of the terms composing an article. All of the models and framework are constructed in Python. The *Gensim* Řehůřek and Sojka (2010) library was used for both the preprocessing of the raw data and the LDA model implementation. The *Spacy* Honnibal and Montani library was used for certain aspects of the text preprocessing such as stop words removal and lemmatization.

2.1 Measuring Uncertainty

In order to evaluate the magnitude of fear and uncertainty embedded in the financial news articles during the *COVID-19* pandemic, a *à la* Baker et al. index was constructed. The idea behind this newspaper-based measurement for policy uncertainty is that information about PU can be enclosed into news articles, however, such articles do not cause uncertainty. They constructed this index by using data from 10 of the most popular US newspapers. The index reflects the monthly ratio of articles related to economic uncertainty and the total number of articles published in the same day. The numerator is the total count of articles that satisfy the following Boolean search over the daily articles: $(economic \vee economy) \wedge (uncertainty \vee uncertain) \wedge (congress \vee deficit \vee FederalReserve \vee legislation \vee regulation \vee whitehouse)$. This triplet filters article that belong to each of the category above: *economy*, *uncertainty* and *policy*. Moreover, in order to ensure that the output is consistent a human auditing, in which people manually reviewed samples of the articles in order to corroborate the fairness of their count, was implemented.

$$EPU_t = \frac{EPUcount_t}{TOTALcount_t} \tag{1}$$

Instead of focusing on sources covering diverse topics, the index proposed in this paper is based solely on Financial Times articles for the *English* news and *ilSole24Ore* for the Italian newspaper. Given the fact that these sources are specifically financial news, I was able to simplify the conditions in the search to construct the numerator. In fact, the nature of these journals is naturally a filter for economic news. Although this may introduce some noise in the data, it ensures that all the articles have an economic or financial background. The queries for the English and Italian based newspapers are as follows:

British search terms:
$(fear \vee risk \vee uncertain \vee uncertainty \vee risky) \wedge (policy \vee tax \vee spending \vee regulation \vee budget \vee deficit \vee parliament \vee bankofengland)$.

Italian search terms:
$(rischio \vee paura \vee incertezza) \wedge (politicamonetaria \vee recessione \vee bancacentrale \vee parlamento \vee bancaitalia \vee legge \vee inflazione \vee deficit \vee tassa)$

Based on the same framework I have also calculated an Economy Uncertainty (EU) index where I have not included policy related term. The second index represents the trend of EU which is embedded in the news. By comparing the two indices one can identify and differentiate periods and related events which may or may not involve economic uncertainty. For example, a spike in the EU index which is not present in the EPU index represents an event that brings some risk or uncertainty in the economy but has not concerned policy makers or regulators. The monthly time-series obtained covers a 13-year period from January 2007 to July 2020. In order to compare the change in time of the two indices the series is standardised to mean 0 and standard deviation 1. Periods with values greater than 0 indicate above average news discussing the terms in the search query. After constructing an overview of the EPU and EU indices, the above methodology is repeated for daily Economic Policy Uncertainty and Economic Uncertainty time series starting from January 2020. This focuses on how Italy and the UK have addressed the different aspects of the pandemic which characterised the first half of 2020. One of the issues with policy uncertainty metric is that it does not take into consideration the notion of *directionality*. In other words, it does not distinguish if the article that has passed the search criteria refers to an increase or decrease in uncertainty. Baker et al. found that newspaper editors and journalists talk about uncertainty when it is high and rising but not when it declines. Moreover, they have shown that human auditing has yielded results showing that only around 5% of the articles analyzed are primarily in the context of policy uncertainty falling.

2.2 Latent Dirichlet Allocation

For the purpose of further investigation, critical periods were identified through the results of the previous indices and the *COVID-19* crisis. Then, an analysis on each article at raw-text level is conducted in order to understand what topics and events where discussed. To extract information from raw text, the Latent Dirichlet Allocation (LDA) was used Blei et al.. The idea behind this probabilistic framework for topic modeling is that articles exhibit multiple topics. Topic is defined as a group of terms belonging to the same theme. Formally, it is defined as a probability distribution of a fixed number of topics K over terms in the $N-1$ simplex, where N represents the number of words in all of the documents. It is important to point out that the population of topics stays the same over all the articles. However, the proportion of topics changes for each article. This characteristic places this model in what statisticians refer to as mixed-membership models, where the mixture proportion changes from article to article but the mixture components stay the same. Mathematically it can be represented as K probabilistic vectors: $\beta_k \in \Delta^{N-1}$ and components are fixed across the entire article collection. Each document, on the other hand has its own distribution over topics: δ_d. Blei et al. in their original LDA paper illustrated the algorithm with the graphical representation reported in Figure 1.

The nodes of Figure1 represent random variables that could be observed (shaded node) or unobserved (white node) and in return can be estimated through the analysis of observed variables. The edges (arrows) are dependencies between random variables represented by the nodes: topic assignment is dependent on document topic distribution, and the observed word in each document is dependent on $z_{d,n}$ and all the topics. The extremes nodes correspond to the hyper-parameters, η and α, for the prior distribution of β_k and δ_d, respectively. The rectangles represent replication variables. The D plate represents the number of articles in the collection, N is the number of unique words and K is the number of topics. This graphical model defines the factorization of the joint probability distribution of the random variables represented in the graph. Additionally, it defines the probability distribution,

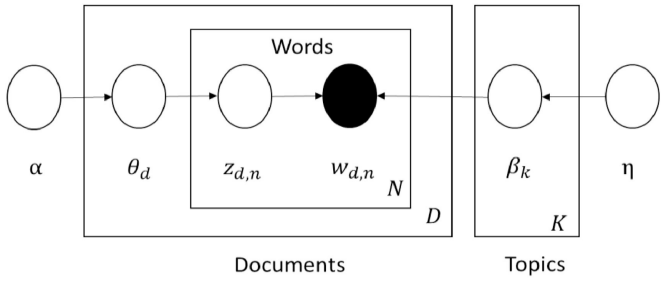


Figure 1: Latent Dirichlet Allocation for topic modelling

posterior distribution, of the hidden structure given the observed article w: $p(\theta, z, \beta|w)$. There are different ways to calculate this posterior distribution. For the purpose of this paper the *Mallet Gensim LDA wrapper* McCallum has been chosen which uses the collapsed Gibbs sampling technique to approximate the posterior distribution as proposed by Steyvers, Smyth, Rosen-Zvi and Griffiths (2004). Several exploratory models were generated in order to find the optimal number of topics. Afterwards, their coherence value was computed and compared in order to find the right model. Ideally, the aim is to choose the number of topics before the plot of the coherence values flattens. After that point, an extra topic is not adding any more useful information.

2.3 Data Preprocessing

Text preprocessing is a crucial step in the analysis of any Natural Language Processing (NLP) problem. It has the objective to reduce the dimension of the dataset by only selecting the tokens which convey some sort of information. The preprocessing of the texts comprises of the following steps (1) lower-casing and special characters removal, (2) tokenization of the articles into terms, (3) stop-words removal, (4) construction of linguistic roots for each term using lemmatization, (5) multi-word tokens creation. Using Regular Expression (RegEx) commands all articles are transformed into lower case and the punctuation is stripped away. In step 2 each article is split into individual elements of interest (terms). Stop-word removal (step 3) is an important and delicate part, it consists of removing words such as articles and prepositions which do not convey any meaning and their presence increases the size of the dataset making it more challenging to extract information. A word can assume different forms within written text, for example, *better* and *best* are transformation of the word *good*, which is the lemma of those words. Lemmatization (step 4) consists of creating the linguistic root of terms. The last step generates multi-word tokens also known as bi-grams and tri-grams. This technique creates a token made up of a sequence of two or three terms which individually would have less meaning during the text analysis. To have the data ready for the topic modeling the preprocessed articles are turned into the document-term matrix. A visual representation of the preprocessing steps is illustrated in Figure2.

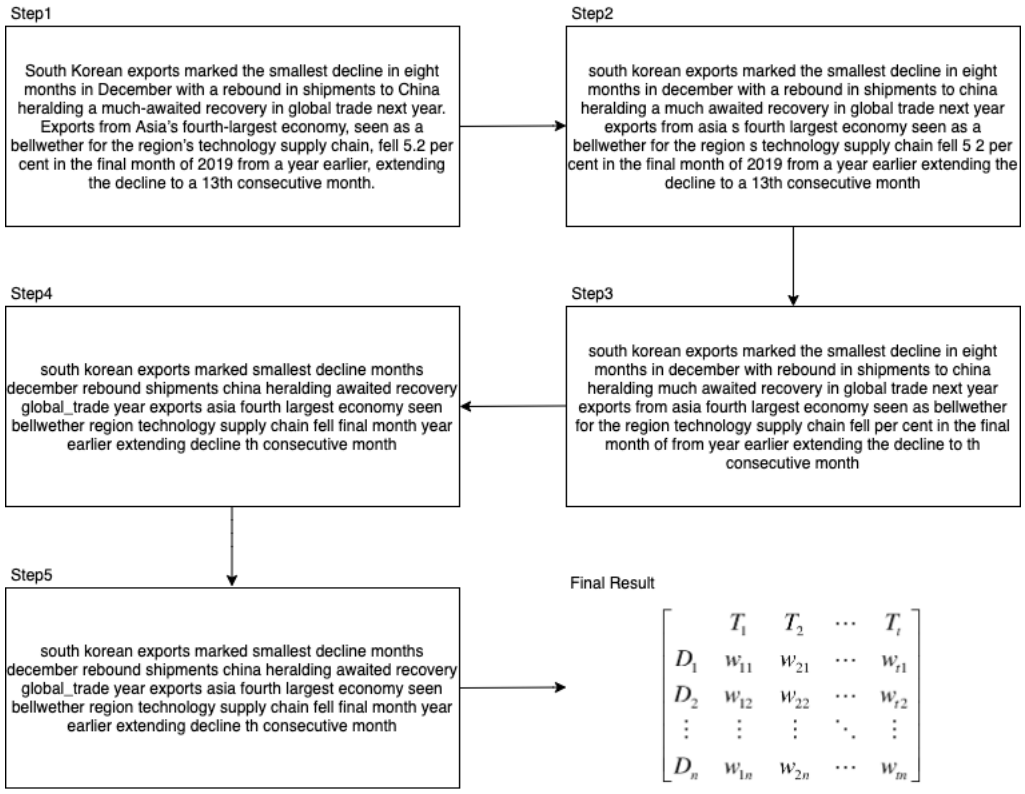


Figure 2: Text Preprocessing

3 Economic Policy Index

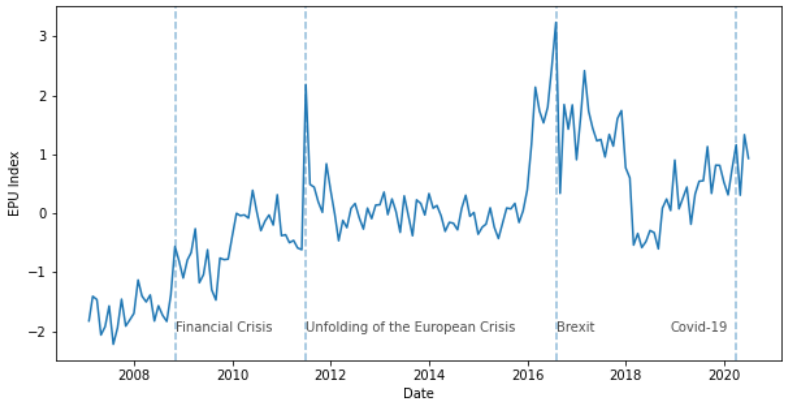
This section of the research implements the construction and analysis of the uncertainty indices for Italy and the United Kingdom as described above. It also proposes a comparison between the two indices and their change over time. Moreover, it outlines stylized facts for both domestic and global events.

3.1 United Kingdom indices

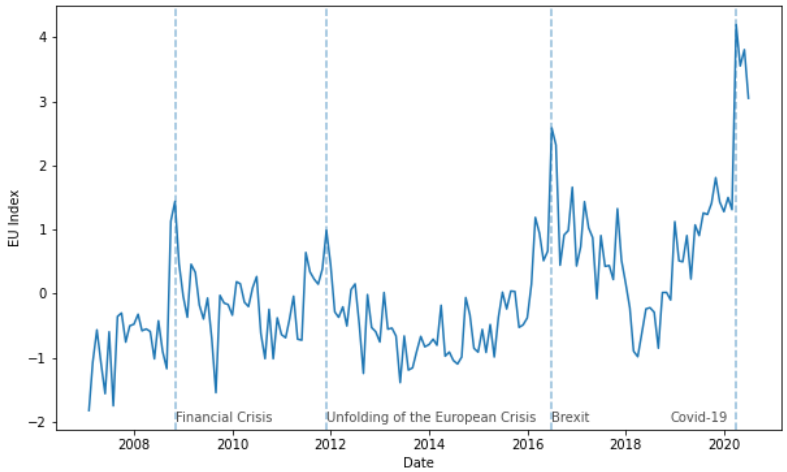
To construct the indices for the United Kingdom daily articles counts from the *Financial Times* were collected. For the numerator the count represents the number of articles that pass the search restriction that have been explained earlier. The results can be observed in Figure 3. The event that has concerned British policy makers and regulators can be clearly distinguished through the analysis of the two indices. In panel a) of Figure 3 I have pointed out the two occurrence that correspond to spikes in the EPU plot. These spikes line up with the unfolding of the European Sovereign Crisis in 2011 and the Brexit Referendum in 2016.

From 2011, the Eurozone witnessed the worsening of the conditions regarding the sovereign debt crisis. Despite budget deficits shrinking across EU members, the concerns of another banking crisis rose. The effects of stimulus policies implemented to bounce back from the 2008 recession started fading away and the weakest economies began to shake. Particularly, in these years the situation in Greece was at the center of attention for economists and regulators around the world as the fear of the Hellenic-republic leaving the European Union was becoming more realistic rather than just a mere threat. Furthermore, the crisis of the most unstable economies affected the stability of countries that experienced an impressive recovery since 2009 and they seemed to fall into a recession period again. For example, it is the case of Germany that after 2011 witnessed a fall in demand in exports. Although not directly involved to the United Kingdom uncertainty increased due to the consequences of global events. In November 2011, the UK's unemployment level reached its highest peak since 1994 and Bank of England governor at the time, Sir Mervyn King, warned that the risks arising from the European crisis could potentially put UK's economy at *'great risk'*. Furthermore, on the other side of the Atlantic, in the United States, many concerns arose around a new fiscal policy and the healthcare system. Ultimately, the leadership transition in China contributed to economic turmoil on a global level. All of these events contributed to the spike in the EPU index in 2011. The second event that had the index spike, setting

the highest jump ever and fluctuating around above average levels from almost two years was the Brexit referendum in June 2016. In the Bank of England 2019 Monetary Policy Report the authors cover an entire paragraph on the effect of Brexit and specifically the impact it had on the uncertainty and its consequences on the country. It stands out that it affected both economies, at an industrial level as well as household level. The most alarming point at the time was the uncertainty regarding the change in the relationships between the UK and its trading partners. Results from the Decision Maker Plan (DMP) survey, show that Brexit caused pessimism about the economy to increase in both households and firms with over 55% of companies placing Brexit amongst the top three reasons for risk regarding the future. The 2008 financial crisis caused only a small increase in the value of the index, however it is still a major spike if compared to previous levels of the EPU. A possible reason



(a) Financial Times EPU Index



(b) Financial Times EU Index

Figure 3: EU and EPU United Kingdom

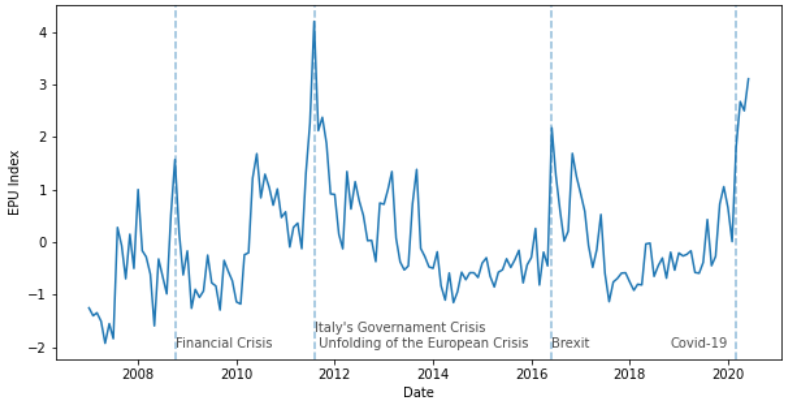
that one of the world's biggest financial crisis has apparently so little impact lies in the availability of the data. In fact, the content of the data downloaded previously 2007 was both scarce and sparse. Including those in the analysis would introduce more noise than information and for this reason I decided to start the time-series on January 2007.

Panel b) of Figure 3 represents the trend of the Economic Uncertainty index. The main difference from the EPU index is that it includes all events that might generate uncertainty without imposing a restriction on policy or regulations. The comparison between the two indices clear indicates which events were most related to policy uncertainty and which were linked to an increase in economic risk in the most various sectors. It is clear at first glance that the financial crisis of 2007-2008 had a greater impact on larger scale as the peak on the graph is higher for the EU index. Considering the variety of sectors and number of people who suffered from this event a large overall risk could be expected, however an *a priori* analysis would suggest a large involvement of governments and policy maker that would cause the EPU index to spike. On the other hand, the result for the 2011 European Sovereign Debt crisis lines up with the expectation. The higher spike in the EPU with respect to the EU index corroborates the nature of the crisis. Debt crisis usually involves the collaboration between governments, monetary policy actions, regulators and central banks. The highest values for both indices occurred during the Brexit Referendum in 2016. Once again, this clearly displays the magnitude of Brexit and its impact not only on British economy but rather on a global level.

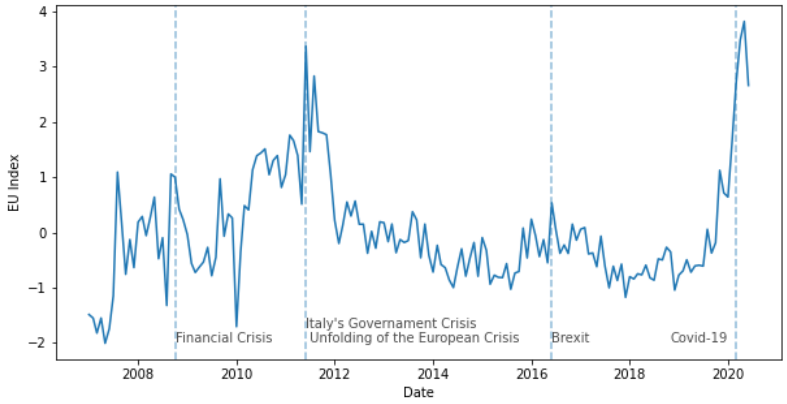
3.2 Italy indices

The same EU and EPU indices were replicated based on the articles from *ilSole24Ore* following the procedure described in section 2.1 in relation to applying the appropriate search query to filter for economic and policy uncertainty. The results are graphically reported in Figure4. Moreover, the graphs have been annotated to the same global macro economic events as the UK indices. While the 2007 Financial Crisis, the European Sovereign Bond Recession and Brexit were occurrences that had an impact at the European level, 2011, amongst other things, was a troubled year for Italy. This was mostly caused by the fall of the *Governo Berlusconi* which triggered a series of issues within the Italian system, including the government crisis. Given this knowledge of the historical course of events a deeper effort was put into investigating the aforementioned period. In order to discover the main events behind the surge in political uncertainty and determine the relative concept of uncertainty associated and how it evolves over time, I conducted an analysis at article level. To answer

those questions, for each article, all tokens were extracted in a 10-word window around the following terms: '*rischio*', '*incertezza*' and '*paura*', which are the italian equivalents of *risk*, *uncertainty* and *fear*. To assess the most discussed topic around the above concepts the weekly frequency on the tokens extracted was computed and ordered in relation to the most frequent to least frequent. By analyzing the time series of the obtained frequencies one can have an idea of how the newspaper associated the concepts of risk, uncertainty and fear over time in the analyzed time period. Table 1 represents the 5 most frequent word in each article and how they change over time. In section 3.2.1 there is an overview of the major event and their correlation to the index. The government debt and the politic issues that led to the fall of the government in late 2011 are clearly visualized in both indices. However, the highest peak in EPU lines up with the expectations as this period included a heavy involvement of government policy makers and central banks. The Brexit Referendum did not cause major



(a) ilSole24Ore EPU Index



(b) ilSole24Ore EU Index

Figure 4: EU and EPU Italy

uncertainty situations in the overall economy in Italy, however there is a jump in the EPU index. Given that Italy was among one of the first countries that was severely hit by the *Covid-19* pandemic it took both the Italian citizens and the government by surprise. With the status of emergency declared by Prime Minister Giuseppe Conte the societal level of uncertainty was rising. For this, it is not surprising that both indices spiked reaching levels close to the 2011-2012 crisis.

Date	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
2010-01	azienda	referendum	confino	crisi	mettere
2010-02	referendum	tragedia	conti_pubblici	truccare	grecia
2010-04	contare	diverso	variabile	finanziario	reale
2010-05	grecia_irlanda	sostenere	wall_street_journal	articolare	prendere
2010-06	mercato	crisi	europeo	crescita	rischiare
2010-07	dovere	arrivare	crescita	potere	aumentare
2010-08	potere	crisi	riprendere	italiano	mercato
2010-09	politico	crisi	presidente	ce	riprendere
2010-10	crisi	mercato	riprendere	potere	restare
2010-11	mercato	crisi	presidente	crescita	potere
2010-12	europeo	restare	dovere	cambiare	aumentare
2011-01	politico	crisi	litalia	ce	crescita
2011-02	petrolio	prezzo	crisi	politico	investimento
2011-03	ce	prezzo	dovere	riprendere	cinese
2011-04	nave	crisi	ce	correre	mettere
2011-05	crisi	impianto	rimanere	prezzo	italia
2011-06	mercato	portare	livellare	prezzo	giornata
2011-07	vedere	economia	mercato	debito	litalia
2011-08	mercato	crisi	cavallo	elevato	crollare
2011-09	crescita	portare	mercato	potere	aggiungere
2011-10	recessione	leconomia	presidente	crisi	possibile
2011-11	capitale	prestito	banca	impresa	finanziamento
2011-12	mercato	capitale	credito	dovere	investimento
2012-01	mercato	impresa	dovere	rischiare	difficolta
2012-02	crisi	impresa	credito	spiegare	mercato
2012-03	debito	litalia	costare	italiano	finanziario
2012-04	mercato	crescita	aumentare	potere	globale
2012-05	dovere	grecia	restare	economico	aumentare
2012-06	arrivare	europeo	perdere	politico	volere

Table 1: Most frequent words associated to the concepts of *rischio*, *incertezza* and *paura* in Italy

3.2.1 Stylized Facts and major events, 2010-2012

January – February 2010: As shown by the table, one of the most frequent words during these months is *referendum*. It refers to the referendum organised by one of the major Italian company, Fiat, which was essential on the economy's recovery for one of the biggest factories in Italy. Its successful result led the EPU index to drop to a very low value.
March - June 2010: The index surged with respect to the previous months. This was due to the Greek government-debt crisis, particularly in May, the European Union along with the International Monetary Fund (IMF), decided to loan 110 billions of euros to Greece in order to avoid bankruptcy. Moreover, in the same months Ireland experienced the Irish property bubble, with a substantial decrease in house prices and loans approval.
July 2010 – January 2011: The end of 2010 saw a positive trend of both American and European markets. Significant measures that drove positivity and improved the global market through the 787 billion dollars of stimulus package approved by Obama's government and the European Financial Stability Facility (EFSF).
February - August 2011: In these months the EPU skyrocketed. This significant rise reflected a very difficult period for the Italian economy, in fact the value of the spread *BTP-BUND* substantially increased (from approximately 180 points to 390 points). Considering this financial problem and some internal political issues, the EPU reached its maximum value in August.
September - October 2011: Following some political manoeuvres aiming at the stabilization of the spread the EPU decreased with respect to the record levels reached during the peak of August. However, following the debt downgrade by Standard & Poor's, the Italian financial market severely suffered with an exponential increase in the Credit Default Swap (CDS). Moreover, the French and German prime ministers ordered Silvio Berlusconi, Italian prime minister at the time, to take measures in order to reduce debt. The unsuccessful attempt resulted in the fall of the government which is reflected in the spike of the index.
November 2011 - March 2012: In early November 2011 the *Presidente della Repubblica Italiana* decided to assign the leadership of the country to a government made by economists, led by Mario Monti. The objective of this mandate was to reduce the spread and public deficit by triggering an economic recovery.

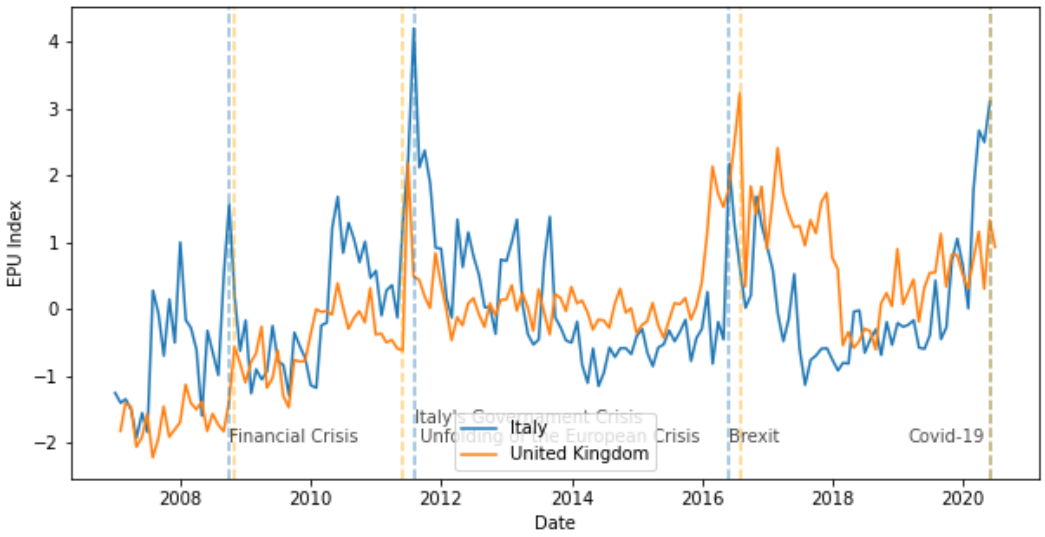
3.3 Indices Comparison

More interesting information can be drawn from the comparison between the two EPU indices. Figure 5 reports the monthly comparison between the two countries (Panel a) and the 12-month rolling window correlation (Panel b). From the first plot it is possible to assess the extent to which regulators and policy makers were involved over time in each country. Panel b) depicts the 12-month rolling window correlation between the two uncertainty measurements and how their *Pearson* correlation coefficient evolves over time. Panel a) suggests that during the 2007 - 2008 Financial Crisis Italy was more affected on the political front as represented by a higher peak in the graph. The trend over the following year up to the European Sovereign Debt crisis lines up with what has been discussed so far. Both countries have been affected by the economic and political turmoil that characterized those years and it is clearly visible. However, as stated in section 3.2.1 Italy also suffered from a government crisis. This is reflected in the index huge surge in late 2011, a much steeper increase for Italy with respect to the UK. The consistency continues over time, and is evident during the Brexit referendum when the indices spike again. This event had a great impact on an international level, however it is not surprising that the peak is sensibly higher in the United Kingdom. Panel b) reports the change in the correlation over time. In particular, it plots the 12-month rolling window *Pearson* correlation coefficient. The interesting fact about this particular graph is that it spikes when the indices are correlated. With peaks reaching almost over 80%, the time periods that correspond to local *maxima* in this plot suggest a strong correlation between the two indices. This implies that both the *Financial Times* and *ilSole24Ore* discussed global macroeconomic events with a strong impact on different policies. These macroeconomics events are summarized in Table 2.

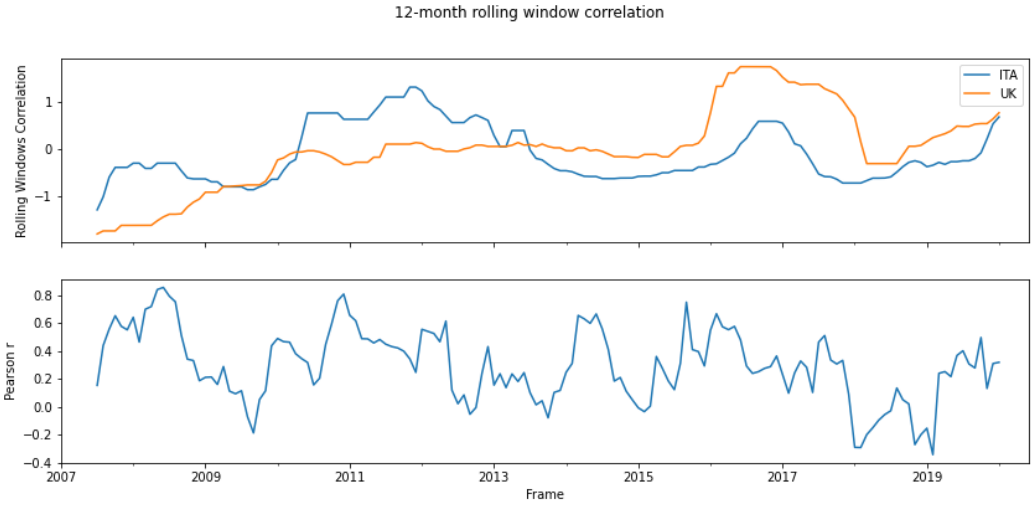
The *COVID-19* has undoubtedly affected the global economy and in general every aspect of society. The remainder of this paper will focus on different aspects of the pandemic. First, the focus will be on constructing daily Economic Uncertainty and Economic Policy Uncertainty indices followed by the analysis applied to the topic modelling techniques describe in the previous section.

PERIOD	EVENT
2008	Financial Crisis
2010-2011	European debt crisis
2014	Oil Prices Crash
2016	Brexit

Table 2: Summary of macroeconomic events affect both countries as represented by the correlation between the two indices



(a) Monthly Comparison



(b) 12-month correlation

Figure 5: Italy and UK EPU comparison

4 COVID-19 analysis

Since the trigger of the global COVID-19 pandemic, the virus has engendered many consequences, beyond its impact on the human health. Along with its exponential spread, it also generated a huge surge in uncertainty which led to many disruptions in our daily life aspects. Many unsettled questions emerged, involving various fields amongst healthcare, epidemiology, economy, public spaces, mental health and education for example. It also led to the reassessment of existing policies to serve the state of emergency where healthcare system got to the verge of saturation. The unawareness of the virus' onset and emergence left governments rushing to make the better decisions with the least effects on the population's health and/or the country's economy. As each government proceeded differently, many plans of actions were adapted to fast adjust to the situation. Some countries prioritized the people and their health first by imposing a set of social policies such as social distancing and total lock-down in order to contain the exponential spread of the virus and flatten the curve. Other countries focused foremost on the economic and financial consequences of the pandemic by involving policy makers and regulators in order to avoid the expected recessions in their economy. This section of the paper attempts to outline and understand the different approaches that both Italy and the United Kingdom have taken over time to engage with the virus. First, I will be using the text based measures already implemented in other sections of the research to comprehend the areas of uncertainty that the leaders of these countries had to deal with along with the biggest questions and concerns. Second, I will be making use of the Blei et al. Latent Dirichlet Allocation model to extract the most discussed topics in the *Financial Times*, on one hand, and *ilSole24Ore*, on the other, in order to evaluate where the newspaper stemmed their attention.

4.1 Economic Uncertainty Analysis

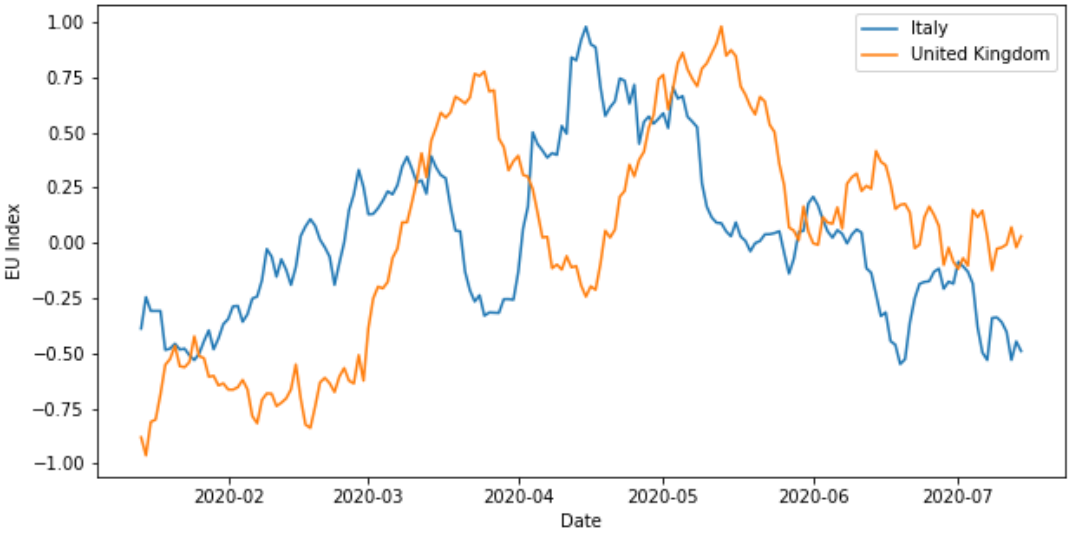
The Economic Uncertainty index is the metric I use in this paper to assess the level of uncertainty not restricted to any specific area. In this section of the research I will use the EU index to analyze how the level of general economic uncertainty evolved in the time of the pandemic. Among the chaos and confusion triggered by the pandemic governments had to rapidly adjust and organize the country to best deal with the situation. Italy was one of the first severely hit countries outside mainland China. Despite a few cases registering in early 2020 the matter started to preoccupy with the appearance of the first hotbeds in mid February in the southern part of the region Lombardy. Epidemiologists and mathematicians

who have been crunching the number of the pandemic highlighted a 2-week gap between Italy and other European regions. The first precaution taken by the government including closing all schools, universities, sports events and leisure activity such as movies and theaters occurred at the beginning of March with the establishment of the lock-down on the entire peninsula on March 9th 2020. It was clear from the beginning that Italy's method with dealing with the virus gave priority to containing the virus in order to ensure that the capacity of the healthcare system would not collapse due to the increasing number of patients requiring treatment in the Intensive Care Unit. On the other hand, UK's approach preferred to avoid drastic situation that could harm the economic situation. However, under pressure of scientists and the public opinion, Prime Minister Boris Johnson ordered a first lock-down on March 23rd, two weeks after Italy. Figure 6 represents the EU and EPU indices for both Italy and the United Kingdom from January 2020 to mid July 2020. The series has been normalized and standardized to mean 0 and standard deviation 1 and plotted its 14-day rolling average in order to reduce noise and extract a more significant signal from the series. It is important to point out that because of these adjustment the daily plots in Figure 6 have been shifted forward with respect to the timeline represented on the *x axis*. The index lines up with events that *ex ante* could be the cause of economic uncertainty. For example it is clearly visible the 2-week gap mentioned above. In more detail, the blue line (Italy) has a first spike in early March corresponding to the first reaction in Italy to the virus during that time. The first peak of the orange line (UK) occurred roughly two weeks later at the end of March 2020, indicating consistency between the index and real events. Over the whole time period considered, the two graphs seem to follow the same trend two week apart accurately reflecting the true state of the world. Moving the analysis forward in chronological order the peak of March is followed by a sudden dip in early April for Italy and mid April for the UK. A very plausible explanation for these could be due to fact that these weeks corresponds to the period when cases of corona virus steepened and reached worrying numbers in terms of deaths and rate of infection. The priority and attention shifted completely from protecting the economy to protecting the healthcare system which was at risk of saturation. Shortly after, the indices surged to record levels. Uncertainty around these period increased again due to imminent expected announcement from the governments that would lead to new measures in light of lifting some of the restrictions imposed by the lock-down. These expectations in reality turned into disappointment with the extensions of the limitation further. In Italy, during a press conference on April 10th, Prime Minister Giuseppe Conte announced that Italians would have to wait at least until May 3rd. On the other side of the English Channel the news came a few weeks later, when in early May Boris Johnson addressed the nation saying that this is 'not the time simply to end the lock-down this week'. The Economic Uncertainty index started to gradually decrease in late Spring early Summer

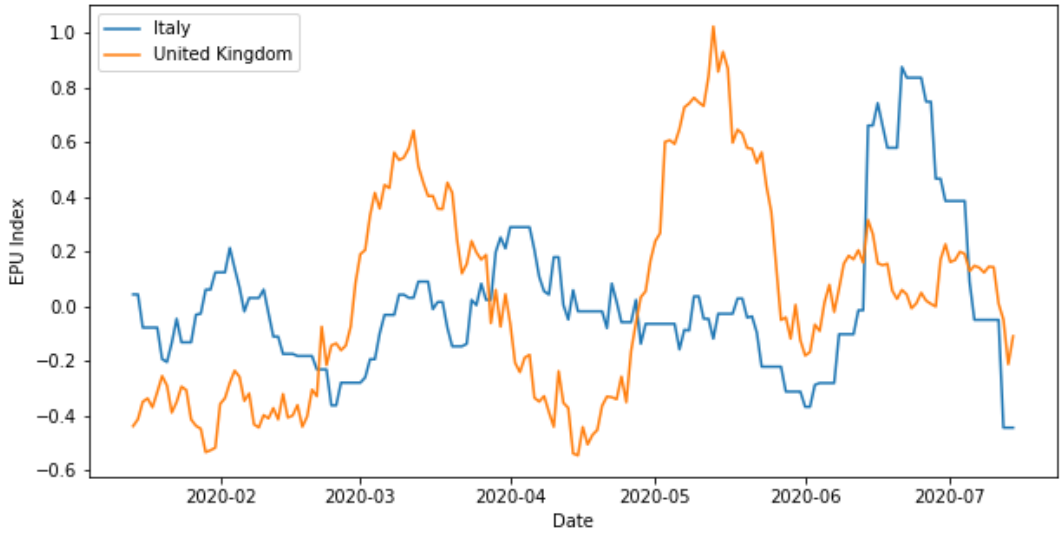
as governments were cautiously lifting the restrictions allowing citizen to restart their life and getting acquainted with the new situation. As society was now more aware on how to deal with the situation, the uncertainty around it began to fade away.

4.2 Economic Policy Uncertainty Analysis

The previous section showed how this new infectious disease has affected the most various aspects of our lives. Many countries however, decided to tackle the issues surrounding economic policy by providing, from the beginning, stimulus packages in order to sustain the domestic economy and avoid its collapse. These measures implied a huge involvement in monetary policy, regulations requiring the help of central banks and economists. Panel b) in Figure 6 shows the Economic Policy Uncertainty index which captures the evolution of policy uncertainty in both countries. In the Italian curve it is possible to see how this country put human health and needs before the status of the economy and waited for the medical situation to stabilize before considering the economy. This is reflected by a shallow trend in the early months of January 2020. In fact, despite a first spike at he beginning of February which can be related to other factors other than the pandemic the index is considerably flat. It is only later in the year that Italy shows uncertainty regarding the economy and policy measures. On the other hand, the United Kingdom has tried since the beginning to tackle the tragic consequences the pandemic was expected to bring on the economy. The uncertainty sharply built up through the weeks leading to the first week of March up until the 12th of March, the same day the World Health Organization declared *corona virus* a global pandemic, Chancellor of the Exchequer Mr Rishi Sunak announced a 12 billion pound emergency package to support British citizens in order to cope with the difficulties brought by the sanitary emergency. Afterwards, in the same manner as the graph in Panel a), the trend drops to minimum levels. In the following months the UK trend skyrocketed towards the end of May when the extension of lock-down measures harshly hit the economy. In these period figures from the Bank of England showed that total debt increased to nearly 28.000 pounds per person. With the restart of the economy and the effect of the policy decisions made so far the confidence in the public increased and uncertainty decreased. The trend of EPU for Italy shows a completely different scenario. The only relevant spike in the graph is towards mid-June however there is not a relevant event that could explain this situation, one explanation could be the fear of a second wave and a second lock-down leading to the fear of a harsher disruption on the economy.



(a) Italy and United Kingdom 14-day rolling window EU index



(b) Italy and United Kingdom 14-day rolling window EPU index

Figure 6: Italy and UK daily indices *COVID-19*

4.3 Latent Dirichlet Allocation

The methodology regarding the Latent Dirichlet Allocation introduced above is now used in order to perform topic extraction on articles published on the digital version of the *Financial Times* and *ilSole24Ore* since January 2020. This section presents the results generated by the model, more precisely, it investigates which topics were mostly discussed during this period and how they evolve over time. *Covid-19* has been the main character for the first part of 2020 and one of the most discussed topics regarding economic uncertainty in newspapers around the world. Baker, Bloom, Davis, Kost, Sammon and Viratyosin (2020) discovered that more then 90% of economic news articles in March 2020 contain *covid*, *coronavirus*, *pandemic* or other term related to infectious diseases. After giving a general overview, the model is applied to pandemic specific articles in order to understand which aspects of the pandemic were most relevant in Italy and the United Kingdom.

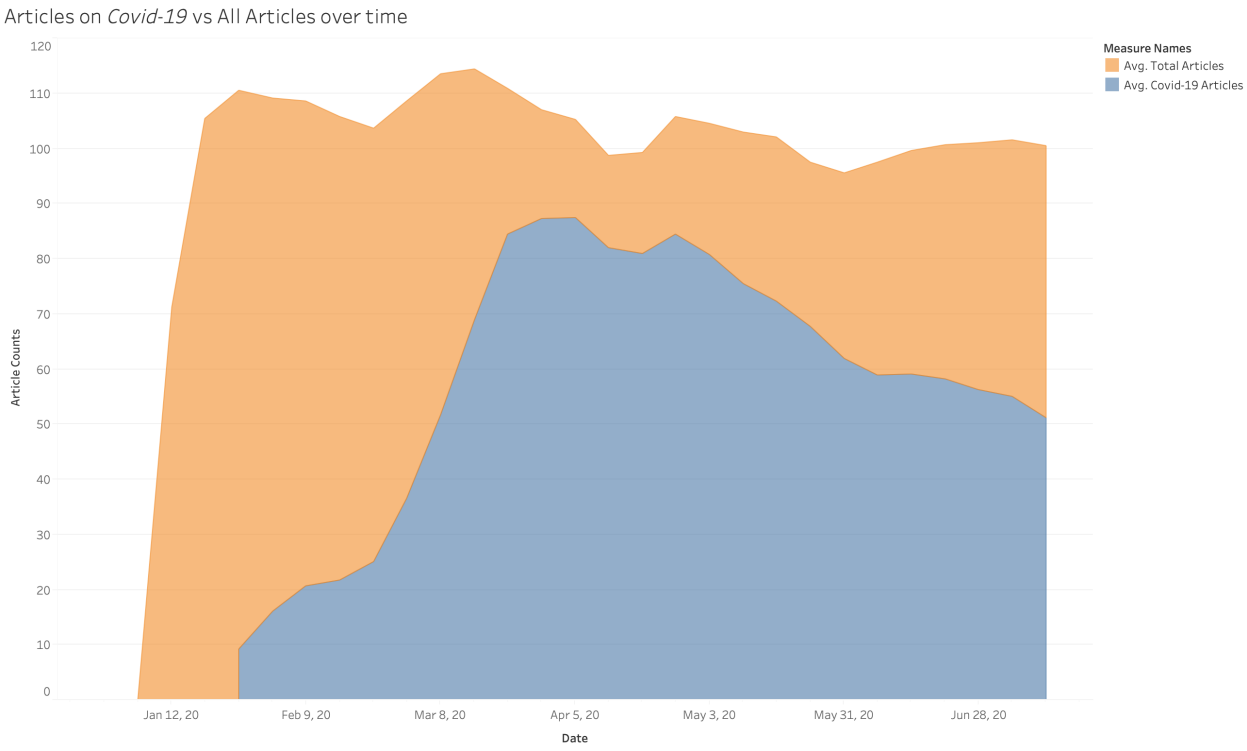
4.4 The Data

In this section I am going to describe the main characteristics of the data used to perform topic extraction on both the *Financial Times* and *ilSole24Ore*. The data regarding Italian news was difficult to acquire, for this reason the dataset obtained is not as big as the dataset for the UK and consequentially more noise is introduced. Nonetheless, significant results were still obtained as shown later on. The final size of both datasets is reported Table3.A more detailed representation of the composition of the dataset is depicted by Figure 7. It represents the 14-day moving average for the number of *Covid-19* related articles against all articles. As expected the proportion spikes in the first month of the year with the ratio $\frac{Covid-19}{Total}$ reaching peaks of over 80% between March and April. From this comparison it also stands out the higher noise in the Italian data. The main input variables for the model are corpus and the dictionary extracted from the articles, as well as the number of topics. Before feeding the data into the LDA algorithm, the texts are transformed following the steps described in 2.3.

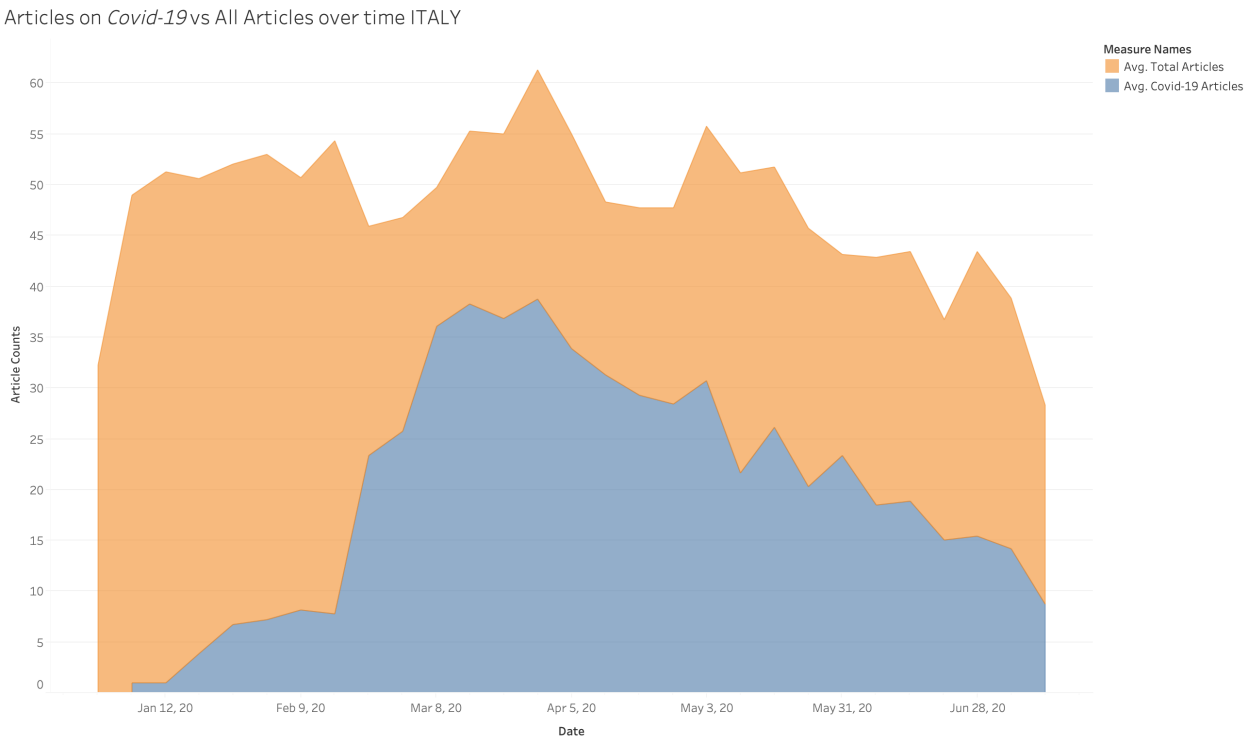
Selecting the optimal number of topics is a delicate task. Several exploratory models were

	United Kingdom	**Italy**
Total Articles	20229	9159
Covid-19 Articles	10268	4148

Table 3: Summary of the size of datasets



(a) All Article vs *Covid-19* Articles UK



(b) All Article vs *Covid-19* Articles ITA

Figure 7: Proportion of pandemic related articles vs all articles from January2020

built with an increasing number of topics. Moreover, in order to assess the best set of parameters for this specific model, the coherence values were used. Existing literature (Newman, Noh, Talley, Karimi and Baldwin (2010) and Syed and Spruit (2017)) proposes different scoring methods meticulously explained, for the purpose of this paper I am adopting the C_v evaluation metrics introduced by Syed and Spruit. The C_v scoring metric is based on four parts: (1) the corpus of the dataset is split into word pairs, (2) for each pair the term pair probabilities is calculated, (3) the likeliness of the sets of words co-occuring with an other pair is evaluated, (4) finally the results are aggregated into a final coherence value. I obtain the coherence scores for 10 different models with a double unit increase from 4 to 26 topics. At each iteration the score value is saved and plotted(Syed and Spruit (2017). The optimal model is selected by picking the number of topic K that is at the end of a steep growth as it is usually an indicator for a model that provides a meaningful output. It might be the case that the coherence score steadily increases with the number of topic. However, picking a too high K might generate granular results dividing each topic into a sub section of micro topic that could cause issues during the interpretation.

4.5 Italy Analysis

Firstly, I construct the topic modeling analysis using all the articles downloaded in order to investigate whether the topics related to the pandemic increase with the passing of time. Figure 8 displays the coherence scores calculated in the manner explained above. From the results, the model with 12 topics is selected. After picking the optimal number of topics the model is retrained with 1000 training iterations in order to reach convergence. The model has now created a word distribution for each of the 12 topics. In order to categorize and analyze how topics change with the passing of time I analyzed over 50 words per topic and manually labeled it according to the content. Table 4 reports the labels and the 5 most frequent word for each topic. The following step consisted of using the model generated previously to extract the topic contribution for each article. Based on this I aggregated the results per month and visualized the time series in Figure 10. The results are clearly in line with the expectations, the first area from the top represent the top share for what has been labelled as *coronavirus*. The topic contribution increased drastically during the first three months of the year and since mid February it became one of the most discussed in Italy. *Economic Burden* and *Economic Uncertainty* followed the same trend corroborating the findings obtained from the indices analysis in the first part of the paper. For the past decade, Italy has been going through a long government crisis and politics have been the center of the attention in the majority of Italian news papers as shown by *Italian Politics*

share at the beginning of the year. As the pandemic hit it witnessed a relative reduction in this topic, suggesting that statesmen united to fight the medical side of the health crisis. However, it did not last long as in mid April the contribution started increasing again reflecting the different ideas and policy proposed by political parties to best re-open and re-start the Italian economy. One last observation worth mentioning in regards to the *education* topic is that schools and universities closed at the beginning of March. This meant that students have not been able to attend lectures since then. The possibility of reopening education institution was a highly discussed matter between politicians in the parliament especially since the lock-down has been lifted. It can be observed in a marginal increase in Figure 10.

Following this general overview, I filtered the articles to extract those related to the pandemic emergency. I refitted multiple LDA models on the new corpus in order to find the optimal one. 9 reports the results. According to this graph 10 topics were chosen, which are displayed in Table 5. The monthly topic visualizations are pictured in 11. What stands out from these results is that the most discussed topics within *covid-19* specific articles are around finance, development and stimulus packages. This is not surprising given that the source is an economic and financial newspaper. The topic shares over time change according to the expectations. At the start of the pandemic there is a big jump in *finance* and topics such as *development* and *production* were not at the center of attention. As time progressed, *stimulus* and *restart*'s contribution increased. Towards the end of the time period analysed, with the reopening of shops and businesses, *development* and *production* were discussed more with respect to previous months.

LABELS	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
Italian Politics	presidente	politico	chiedere	volere	accordare
Italian Manufacturing	produrre	italiano	italia	cliente	azienda
Companies and Development	azienda	sviluppare	investimento	progettare	settore
Education	potere	scuola	studiare	digitale	ricercare
Renewable	sistemare	auto	guidare	offrire	versione
Economic Uncertainty	dovere	crisi	potere	italia	europeo
Economic Burden	euro	previsto	impresa	pagamento	reddito
Arts and Lifestyle	mostrare	museo	artista	evento	opera
Coronavirus	coronavirus	regione	virus	covid	italia
Finance and Markets	mercato	dollaro	calere	aumentare	crescita
Laws and Regulation	potere	attivita	leggere	dovere	diritto
UKN	sapere	volere	storia	venire	potere

Table 4: 5 most frequent words per topic ITA

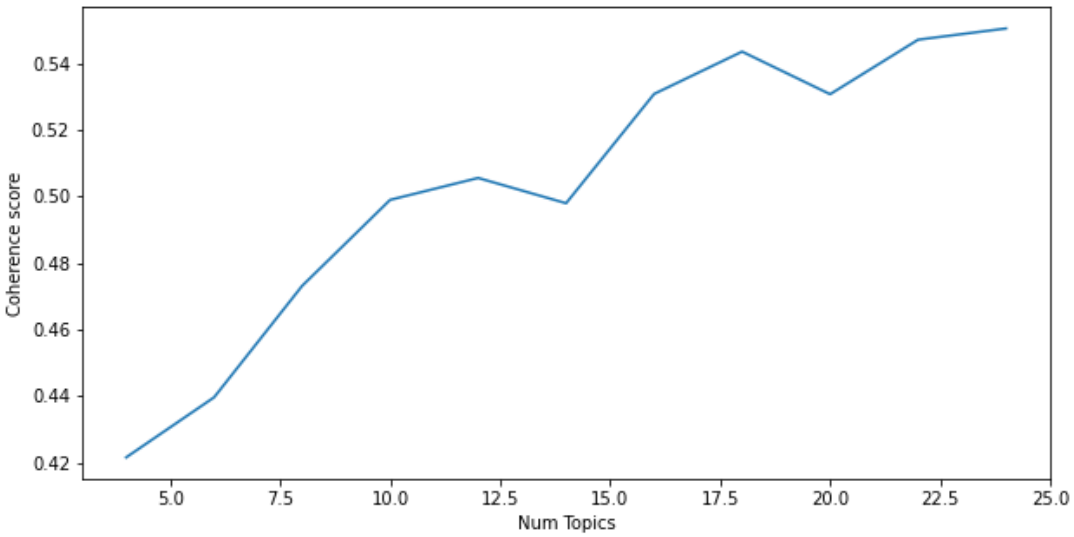


Figure 8: Coherence Score ITA

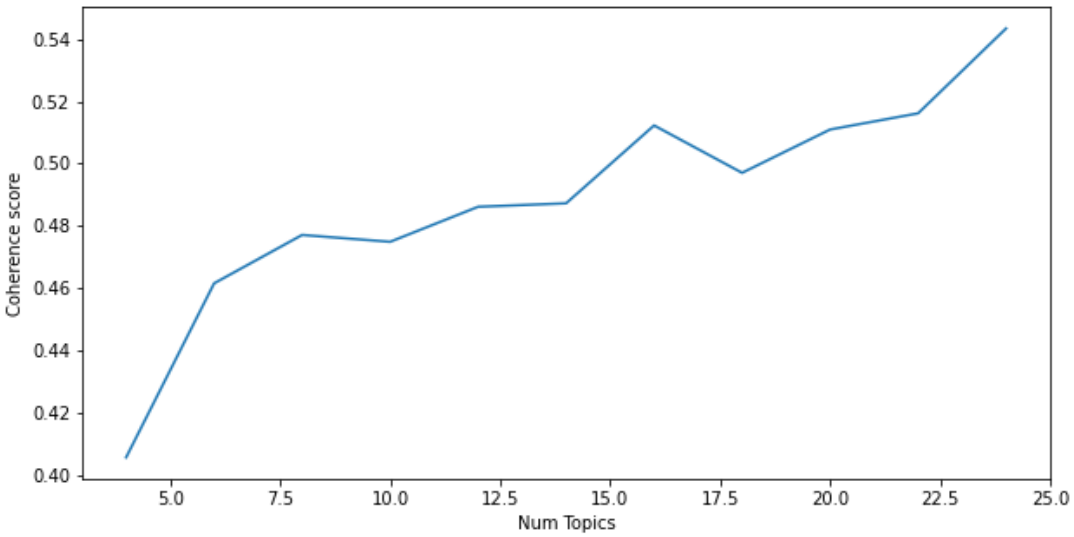


Figure 9: Covid-19 specific Coherence scores ITA

LABELS	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
Production	produrre	azienda	produzione	italiano	euro
Stimulus	euro	impresa	previsto	misura	lavoratore
Development	potere	sistemare	sviluppare	investimento	rete
Outbreak	coronavirus	virus	contagiare	covid	dato
Finance	mercato	dollaro	calere	coronavirus	prezzo
UKN	storia	volere	sapere	vivere	raccontare
Restart	potere	attivita	maggio	sicurezza	aprire
Politics	dovere	presidente	italia	politico	europeo

Table 5: COVID-19 specific 5 most frequent words per topic ITA

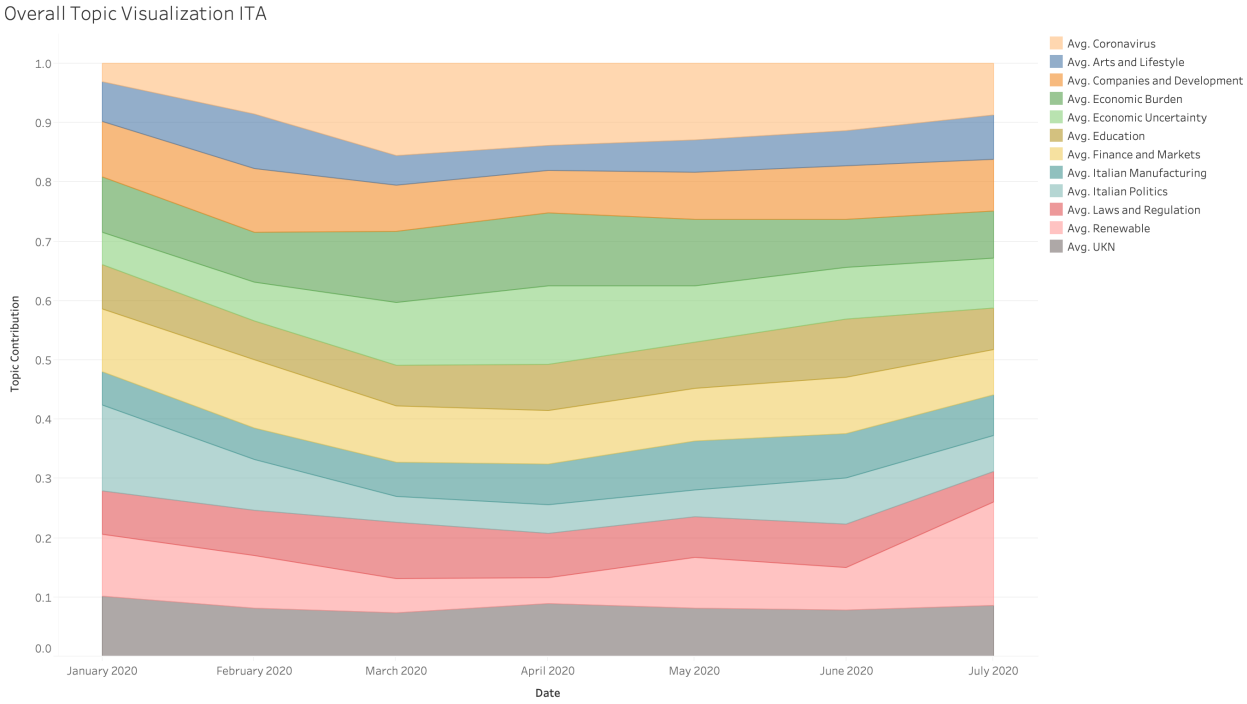


Figure 10: Topic Visualization over time Italy

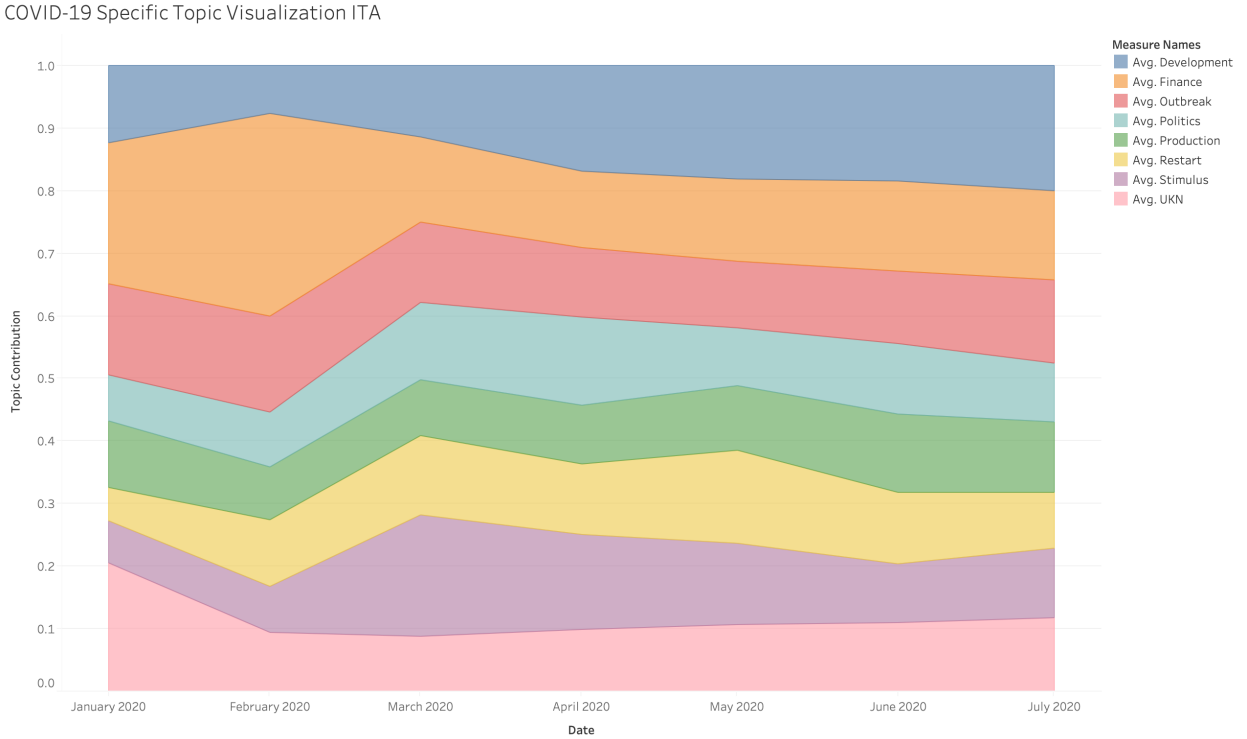


Figure 11: Covid-19 specific Topic Visualization over time Italy

4.6 United Kingdom Analysis

Following the same framework describe in Section 4.5 I conduct topic extraction on the *Financial Times* articles. Collecting the data for the English based newspaper was a much easier process and this allowed me to collect more than double the articles. A richer dataset usually implies that the results are less noisy and more reliable for interpretation. I run exploratory analysis on 11 topics by increasing the number of topics at each iteration. Figure 12 reports the graphical representation of the C_v coherence scores against the number of topics. Using the same *elbow method* as above, I select the number of topics before the line flattens. In this case the model after the 10-topic model causes a significant drop in the graph, hence I select the model previous to the dip. I then re train the model with 1000 passes to reach convergence. Each set of words generated by the model is manually analysed in order to label each topic with the most relevant keyword. The results are reported in Table 7. The model is used to generate the daily topic share and then aggregated by month. 14 visualizes the time series for the monthly topic share. The results line up with the finding presented in the EU and EPU indices section. There is a big contribution from *coronavirus* starting from February 2020 and increasing until April, the period where the virus hit the country the most. In parallel the *economic* topic follows the same trend as *coronavirus*, however at the beginning of the year it was not as small as the first one. This is because the pandemic is a novel discussion which was induced at the start of 2020. The similarity reflects that, as already stated, the United Kingdom was most concerned dealing with the economic side of the crisis. Given the popularity of the *Financial Times* it is also possible to find international topics such as *East Markets* and *US Politics*, however during the time of the pandemic their contribution shrunk. Given the soon to happen American presidential election we can observe a substantial spike in this topic towards the end of the time period.

After filtering for *Covid-19* news articles and running the models I obtain the C_v coherence scores reported in Figure 13 with the 12-topic model representing the optimal model. In Table 7 are reported the manually selected topic labels after analysing the word distribution generated by the model. Finally the graphical representation for the *Covid-19* specific monthly topic share over time is visualized in 15. The results from this last topic extraction are the most consistent with what actually happened during the first half of 2020. The pandemic outbreak is undoubtedly one of the most discussed topics for the first months of the year. Its exponentially fast spreading all over the globe has given rise to concerns and uncertainty around several areas. As we can observe from the visualization sub-topics of this pandemic specific analysis were discussed in the *Financial Times*. Particularly relevant to

the United Kingdom is the sharp increase of the share of topics such as *stimulus*, *finance* and *economy* during the most critical months of the emergency. This emphasizes the importance that the government gave to this aspect of the pandemic. A curious result is the massive increase of *London Life* around May as people were looking forward to lifting of lock-down and restarting their social life.

This sections concludes the technical analysis of the research, in the last part I am going to sum up the results from both the Economic Uncertainty Analysis and the Latent Dirichlet Allocation topic extraction. Finally, I will point out the challenges faced during the research, points for improvement and future works.

LABELS	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
Europe	government	uk	eu	britain	policy
Finance	executive	investor	fund	business	investment
UKN	don	ft	ve	today	wine
US Politics	trump	president	political	russia	vote
East Markets	china	chinese	global	hong_kong	beijing
Tech	technology	research	datum	university	service
London Life	london	film	artist	culture	scene
Business	business	sale	uk	customer	lockdown
Economy	price	economy	investor	rate	debt
Coronavirus	coronavirus	virus	government	linkedin_opens	facebook_opens

Table 6: 5 most frequent words per topic UK

LABELS	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
European Union	eu	uk	europe	government	germany
Politics	government	coronavirus	covid	uk	virus
Outbreak	china	chinese	coronavirus	virus	hong_kong
Oil	price	demand	production	russia	global
Finance and Investment	investor	fund	stock	price	asset
Simulus	business	loan	uk	government	executive
Technology	ft	technology	business	launch	online
Social Media	linkedin_opens	facebook_opens	twitter_opens	coronavirus	report
US Politics	trump	president	political	police	election
Development	sale	business	executive	uk	customer
London Life	london	guest	ve	don	spotify
Economy	economy	rate	government	crisis	economic

Table 7: Covid-19 Specific 5 most frequent words per topic UK

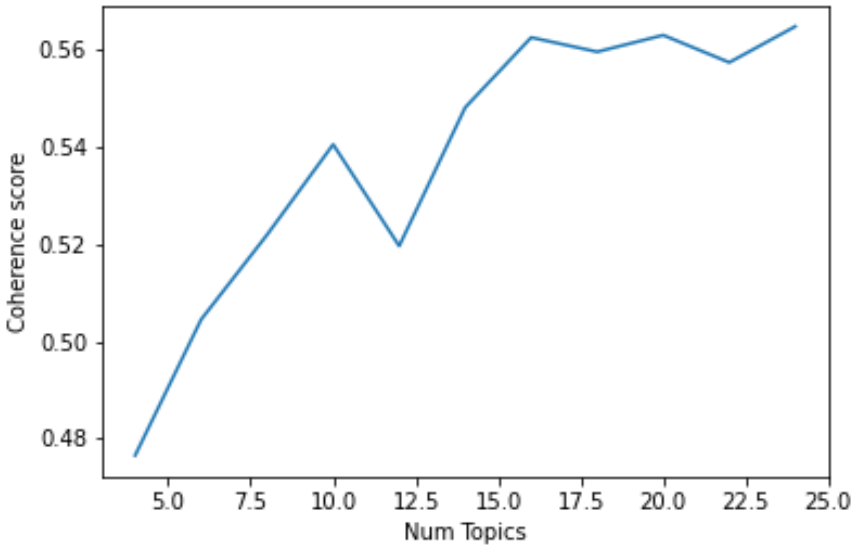


Figure 12: Coherence Score UK

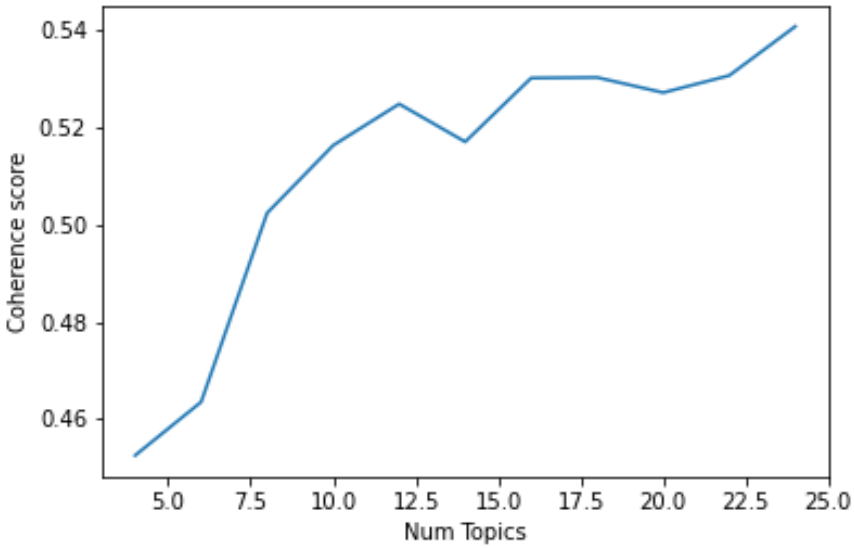


Figure 13: Covid-19 specific Coherence Score UK

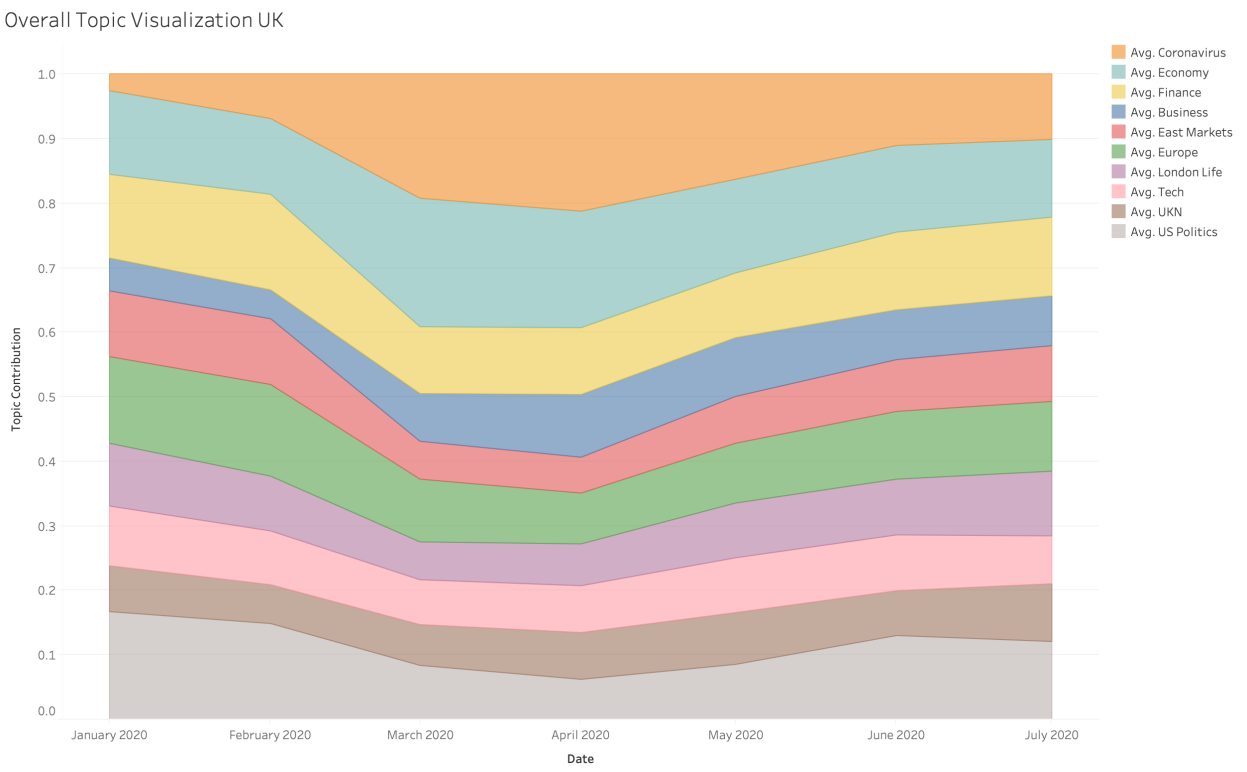


Figure 14: Topic Visualization over time UK

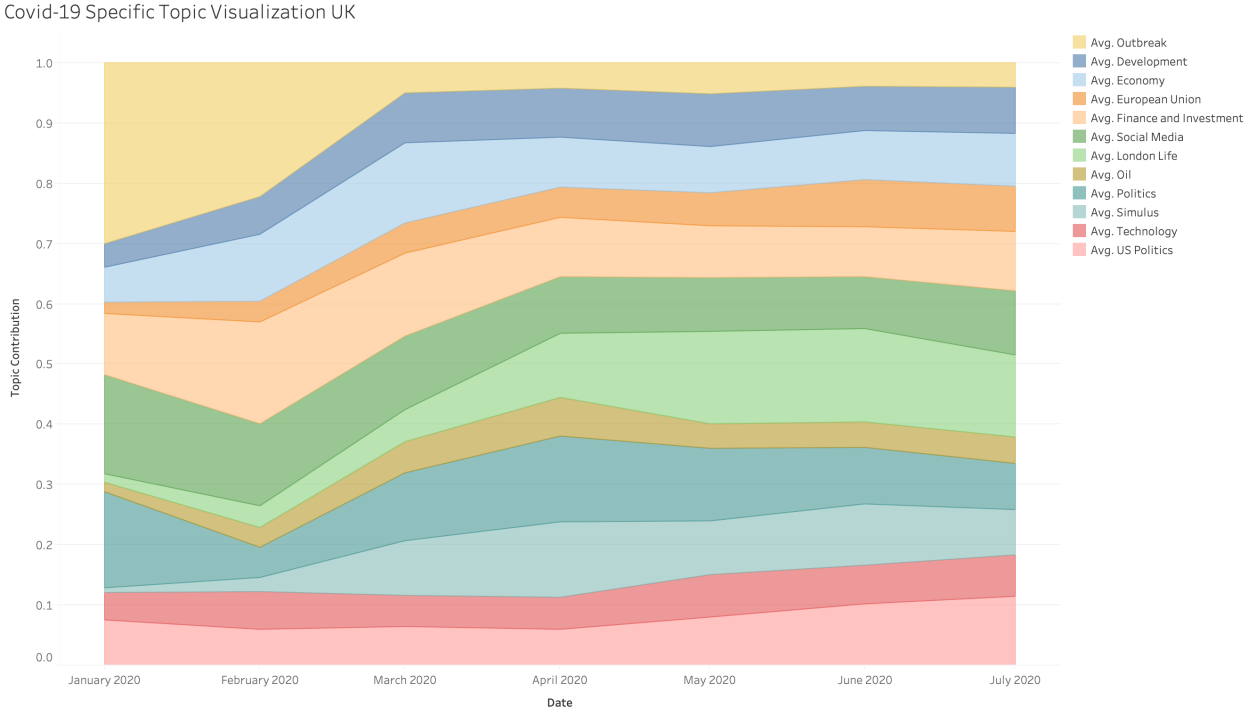


Figure 15: Covid-19 Specific Topic Visualization over time UK

5 Conclusion

In this paper I implemented different frameworks to analyse and assess text based news in order to extract relevant information regarding levels of economic uncertainty. This is a cross country analysis focusing on Italy and the United Kingdom. To retrieve information, the *The Financial Times* and *ilSole24Ore*, two of the most reliable economic newspapers in the United Kingdom and Italy, respectively were used. The research is centered around two time periods. A long term study from 2007 to 2020 investigates the overall trend of economic uncertainty over time, it outlines and assess the methodology used. The main focus is around the *Covid-19* pandemic that hit and radically changed many aspects of our life, including global economy. Using two different newspaper from each country I extrapolate information on how different government dealt with the pandemic and which aspects of it had the biggest attention. For the first part of the research, I have constructed two different text based indices using the methodology proposed by Baker et al. (2016). The first index Economic Uncertainty (EU) is a measure of the overall level of uncertainty, the second one, Economic Policy Uncertainty (EPU) relates to the events that generated policy uncertainty. The result of the comparison between the two countries outlined how the differently approach macro economic events and how these affected the domestic economy. The analysis includes the explanation of the framework used, detailed examples and interpretation of the results. The core section of this paper is the *Covid-19* analysis, where I constructed the daily series for both the EU and EPU index. This clearly showed how the two countries had a different approach in coping with the virus. The results line up with the *a priori* analysis of the events and emerges that the UK was most concerned with policy and regulation to avoid a collapse in the economy, whereas Italy was most concerned in containing the spread of the virus by imposing stricter restrictions and lockdown measures. Furthermore, the EU index trend showed the two week gap in the spreading of the disease which many expert talked about early in the year. Lastly, I used a Latent Dirichlet Allocation (LDA) as proposed by Blei et al. (2003) in order to perform topic extraction on news articles from January 2020 to July 2020. The LDA model was fitted on articles from both countries. For each nation I have extracted a general overview of the topics discussed in the *Financial Times* and *ilSole24Ore* and a pandemic specific where I filtered the articles for terms related to the pandemic. Both results are coherent with the events that characterized the first half of the year 2020. The change over time of the topic shares contribution, the EU index and the EPU index give an accurate picture of the time series of the events during these first 6 months. During this study I have faced several challenges. The main one being the retrieval of the articles. For the indices construction I have used the newspaper's search function to retrieve the daily

counts. The numerator was obtained with a Boolean search with the relevant queries as explained above. The denominator was constructed by search for a neutral term like articles or popular verbs which are likely to appear in every article. The retrieval of the entire articles from January 2020 was more difficult. The *Financial Times* provides and API and I was able to download them. However, this option was not available for the Italian newspaper. In order get those article I set up a web scraper in Python, in this way, however, I was able to get only articles which did not require a subscription to access them. For this reason, the Italian dataset is half the size of the UK one and the analysis resulting from that has more noise. The methodology and results of this research can serve as a starting point for future work. The system proposed can be adjusted to other countries and languages with the appropriate transformation required during the data retrieval process. It is straightforward methodology to apply to other countries in order to get a global overview of the events. It is also not specific to the *Covid-19* pandemic but can be used with any event to analyze how countries are dealing with the circumstance. Ultimately, it can be used by the governments, policy makers and regulators themselves to assess their work and be ready in case similar events occur again.

6 Bibliography

References

Baker, S., Bloom, N., Davis, S., Kost, K., Sammon, M. and Viratyosin, T. (2020), 'The Unprecedented Stock Market Impact of COVID-19', *National Bureau of Economic Research* .

Baker, S. R., Bloom, N. and Davis, S. J. (2016), 'Measuring Economic Policy Uncertainty'.

Balta, N., Fernandez, I. V. and Ruscher, E. (2013), 'Assessing the impact of uncertainty on consumption and investment', *Quarterly Report on the Euro Area (QREA)* **12**(2), 7–16.

Bank of England (2019), Monetary Policy Committee Monetary Policy Report, Technical report.
URL: *www.bankofengland.co.uk/monetary-policy-report/2019/november-2019*

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent Dirichlet allocation', *Journal of Machine Learning Research* **3**(4-5), 993–1022.

Caggiano, G., Castelnuovo, E. and Groshenny, N. (2014), 'Uncertainty shocks and unemployment dynamics in U.S. recessions', *Journal of Monetary Economics* **67**, 78–92.

De Boef, S. and Kellstedt, P. M. (2004), The Political (And Economic) Origins of Consumer Confidence, Technical Report 4.

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

Honnibal, M. and Montani, I. (2017), Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

McCallum, A. K. (2002), MALLET: A Machine Learning for Language Toolkit.
URL: *http://mallet.cs.umass.edu*

Newman, D., Noh, Y., Talley, E., Karimi, S. and Baldwin, T. (2010), Evaluating topic models for digital libraries, *in* 'Proceedings of the ACM International Conference on Digital Libraries', ACM Press, New York, New York, USA, pp. 215–224.
URL: *http://portal.acm.org/citation.cfm?doid=1816123.1816156*

Řehůřek, R. and Sojka, P. (2010), Software Framework for Topic Modelling with Large Corpora, *in* 'Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks', ELRA, Valletta, Malta, pp. 45–50.

Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T. (2004), Probabilistic author-topic models for information discovery, *in* 'KDD-2004 - Proceedings of the Tenth ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining'.

Syed, S. and Spruit, M. (2017), Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation, *in* 'Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017', Vol. 2018-January, Institute of Electrical and Electronics Engineers Inc., pp. 165–174.