Master's Degree Programme
in International Management

Ca' Foscari
University
of Venice

Final Thesis

Brexit and Uncertainty: an empirical and dynamic analysis of an event through Taxonomies and Twitter Data.

Supervisor

Prof. Massimo Warglien

Assistant Supervisor

Prof. Carlo Romano Santagiustina

Graduand

Nicholas Schibuola
Matriculation ID 856997

Academic Year

2018/2019

Index

1. Introduction

In this thesis we analyse data coming from approximately 112,000 UK Tweets posted from May 2018 to May 2019, which contain the term "Uncertainty". Our final aim is to deepen our understanding of how uncertainty is perceived in different geographical areas of the UK, and how people link "Brexit" Uncertainty to aggregate economic and social variables of interest. In order to do so, we use innovative methodologies such as hierarchical taxonomies, which include social and economic variables of interest for this analysis. These taxonomies, which were conceived and developed with C. Santagiustina and M.Warglien for the “Worldwide Uncertainty Observatory” (WUO) of Ca’ Foscari, will be released as open source software. The final aim of the aforementioned project is to “enable researchers, but also an audience of journalists, investors, analysts, managers, students and academics to visualise and analyse the uncertainty of civil society, for multiple geographical areas, in real time” (M.Warglien, 2019).
This work is divided in four parts:

- In the first one, we conduct a comparative analysis between aggregated and disaggregated uncertainty by geographic area, to show differences in the perception of uncertainty in the different parts of the United Kingdom.
- In the second part, we implement the taxonomies to analyse the dataset more in depth and gain insight on how uncertainty is perceived to affect economic variables of interest such as unemployment and inflation. We devote our attention to economic variables co-occurrence matrices, that we analyse at 3 levels of abstraction through network analysis.
- In the third part, we analyse lexical diversity of the dataset by geographical area.
- The last part focuses on the sentiment analysis by geographic area, using two different measurement techniques: classical sentiment analysis and NRC Emotions Lexicon.

1.1 Uncertainty

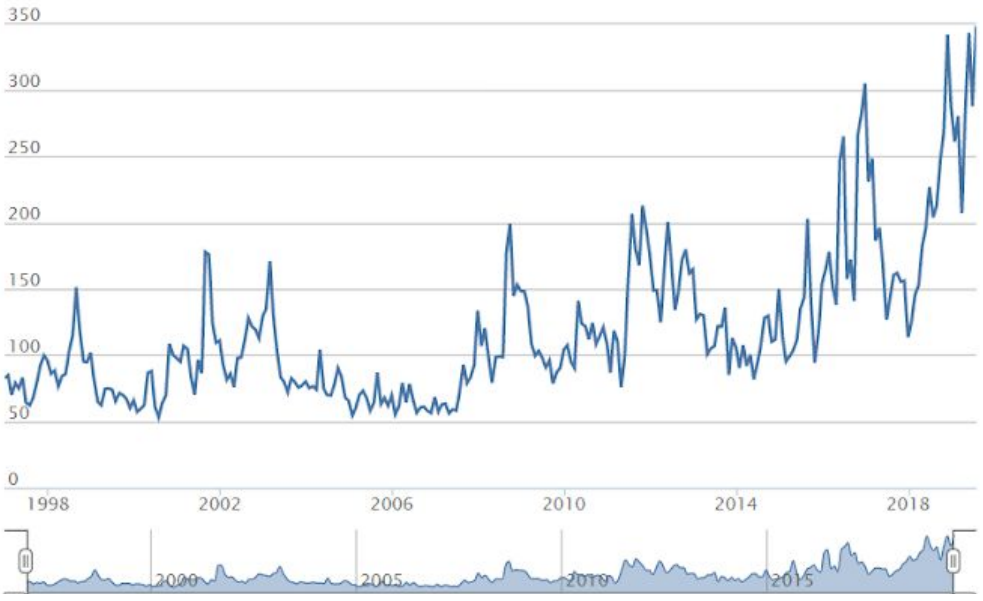
Uncertainty is officially defined "as a situation in which something is not known, or something that is not known or certain" (Cambridge Advanced Learner's Dictionary & Thesaurus). To deepen our understanding of what uncertainty is, we can describe it as "[...] a state in which an agent is reluctant or unwilling to use his belief system for formulating expectations and taking relevant decisions." (C. Santagiustina, 2018).
It is important to study uncertainty and how it is perceived by people because it affects the economy; in particular, households, businesses and financial markets. A report of PwC states that uncertainty could "reduce consumer spending to safeguard against potential future falls in income" and that "this is also associated with an increase in precautionary household savings" (PwC.com, 2017). Concerning the effects of uncertainty on businesses, the report asserts that businesses could need to reduce production, investments and employee salary. Moreover, the financial market undergoes the effects of uncertainty too because "investors require a higher rate of return on their capital through higher risk premia" when there is an uncertain situation. In addition, "at times of uncertainty, capital also tends to flow from riskier to safer asset classes" (PwC.com, 2017). All these effects caused by uncertainty impact the overall economy.
As you can see in figure 1 below, the Economic Policy Uncertainty index, developed by Baker, Bloom and Davis (Economic Policy Uncertainty Website, 2012-2018), has been growing in the last 10 years throughout the world reaching the highest point in 2019. "To measure policy-related economic uncertainty, we construct an index from three types of underlying components. One component quantifies newspaper coverage of policy-related economic uncertainty. A second component reflects the number of federal tax code provisions set to expire in future years. The third component uses disagreement among economic forecasters as a proxy for uncertainty." (S. R. Baker, N. Bloom, and S. J. Davisc, 2016). Given this important information, it becomes useful to find techniques that allow researchers, companies as well as stakeholders to grasp how uncertainty is perceived and, more importantly, linked to economic and policy variables of interest. This is a significant innovation, because it enables researchers to be aware of how uncertainty is structured and perceived to be related to other economic

phenomena.

In this work, we want to understand how uncertainty events in the United Kingdom during the Brexit process are perceived by people. This, in order to know if there is an uncertainty state among people regarding a particular event or phenomenon.

Figure 1: Global Economic Policy Uncertainty Index

(from https://www.policyuncertainty.com/index.html)



1.2 Twitter & Uncertainty

The trend of social media has changed: people don't use them just to talk about themselves but also to share their thoughts about what is happening in the world. This change has already been observed, studied and reported by many researchers. As an example, Richard Rogers, in a paper published in 2013 (Debanalizing Twitter: The Transformation of an Object of Study), states that: "Over the past few years for researchers Twitter has evolved from a phatic and ambient intimacy machine [...] to an event-following and news machine, [...] when the Twitter tagline changed from 'what are you doing?' to 'what's happening?'." For this reason, this freely and publicly accessible platform is an important source of society's point of view related to any kind

of phenomenon. "[...] Twitter can be considered an instrument for collective elicitation and interpretation of global events and expectations, and hence, of the cognitive states explicitly associated to these events and expectations." (C. Santagiustina, 2018).
In fact, Twitter is an online community where every day millions of users interact and share their thoughts and sentiments about events affecting the current world. Therefore, Twitter can help us understand how people perceive everyday events and phenomena such as uncertainty.
But how can we analyse, from a practical point of view, what people think about a certain phenomenon? How can we study uncertainty through some statements that people write on Twitter? Twitter "[...] has settled into a data set, from which researchers have made collections [...]" (Rogers R., 2013). This type of data set has many important properties which allow us to do different significant analysis. For example, the low and fixed number of words that each tweet can contain in order to be posted, make it possible to do text analysis. Thanks to these ones we can find an infinite amount of information that lays behind the various tweets. As a consequence, this source of information is useful to understand uncertainty feelings among people and how people link uncertainty to specific phenomena.

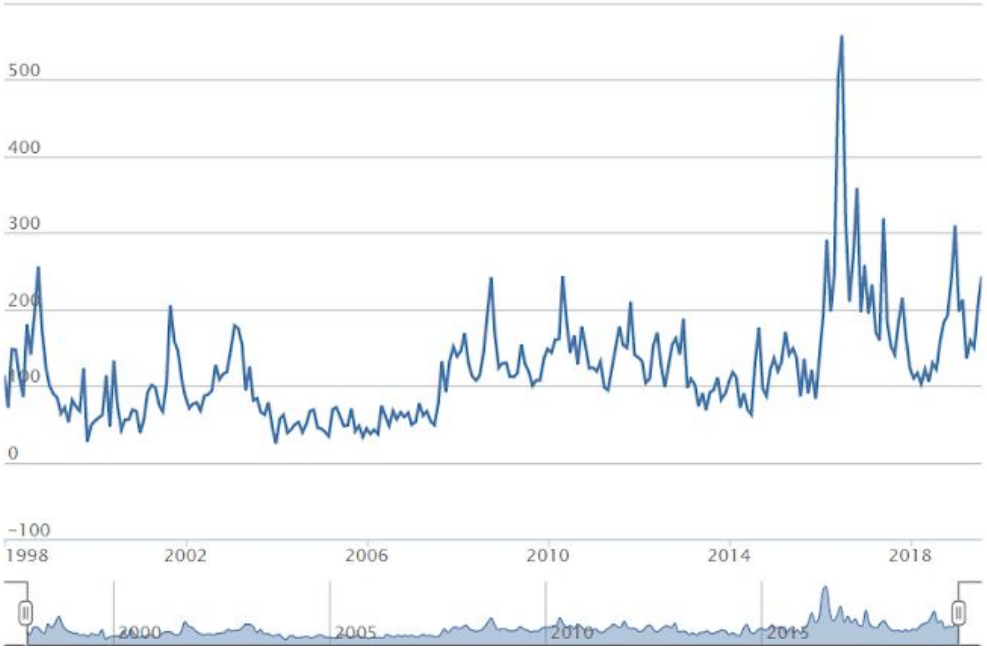
1.3 Brexit

For this thesis we chose to work on a subject that has been in evidence all over Europe and the entire world: Brexit. Brexit is the withdrawal of the UK from the European Union, scheduled after the referendum held on June the 23rd 2016, which revealed that 52% of voters were in favour of a withdrawal from the European Union. The reason why we chose this event is that it is strongly characterized by uncertainty. In fact, as figure 2 below shows us, in 2016 the economic policy uncertainty index has skyrocketed when reaching the highest point ever reached before. Moreover, if we look more carefully at the graph, we can see that this index is growing even right now, in 2019. This trend is reported by many british journals such as "The Economist" (Economist.com, 2019) and its effects on UK economy are reported too. The journal "The Guardian" (Theguardian.com, 2019) explains these effects in an article entitled

"Recession fears grow as UK economy shrinks on back of Brexit chaos". In fact, the article states that gross domestic product shrank by 0.2% between April and June of 2019.

Figure 2: UK Economic Policy Uncertainty Index

(from https://www.policyuncertainty.com/index.html)



1.4 Aim of the Analysis

The final aim of this analysis is quadruple. Firstly, we want to see how uncertainty is perceived through a modern vehicle such as Twitter; in particular, we want to understand how this condition of uncertainty deals with socio-economic variables. Furthermore, we want to better comprehend how uncertainty is linked to other variables such as political actors, businesses, influencers. Then, we want to understand how uncertainty is described. Is it poorly debated from a linguistic point of view? Finally, we want to see what is the sentiment of the people.

2. Method

In this section we define the architecture and methods of this work. Since data analysis is a very broad topic, and there are hundreds of possible analysis to be made for each dataset, we select just the ones that we believe are more coherent with the aim of our thesis. The main software used to analyse the dataset is R (R Core Team, 2019), and its interface RStudio, an integrated development environment for R (RStudio Team, 2018). R is a programming language that allows users to do both simple and deep data analysis and to easily integrate different programming languages so that analysis can be more flexible. In fact, a big chunk of the data pre-processing is done using Perl and its regex (regular expressions) through RStudio. In this way we are able to reach a higher level of precision in preparing the data for the analysis. The majority of the work is done using an open-source library for R, called Quanteda. We choose to operate with this library because it is flexible and, thus, allows to do complex analysis in a few passages and to integrate deep level analysis such as network analysis with our taxonomies. This is the most important part of our work. In fact, the way in which we integrate deep-level analysis such as network analysis and semantic networks with the taxonomies should be considered an important methodological innovation of this work. The results coming from the whole of the analysis are functional to the answer of our research question.
The methodology we use in this thesis is developed to be modular. This aspect allows anyone to reuse the analysis explained throughout this work changing dataset or taxonomies, on the basis of one's own research question. In fact, a further aim of the methodology introduced in this thesis is to provide to the WUO (Web Observatory of Uncertainty) a new instrument, which consists in an always up-to-date website, with real-time data that is already analysed. That is possible using automatization of code, thanks to the use of scripts which ensure real-time update of data. The integration of this operations is possible with the use of R and its flexibility.
We chose to structure our work with not only one but four types of analysis purely because we want to provide a complete overview and answer to our research question. All the packages used in the different phases of the analysis can be found in Appendix.

2.1 Data Extraction (Twitter)

Data Extraction is the first phase of any analysis of data: it consists of the collection of the data that is used in this analysis. Our dataset contains 111,251 Twitter Posts, published from 15th, May 2018 to 28th, May 2019.
The conditions which we pose in order to filter these tweets are three: the Tweet has to contain the term "Uncertainty", has to be posted by a user that explicitly declares as location a place within the UK, and must be written in english. In order to extract data, we use the Stream endpoint of the Twitter API (https://developer.twitter.com/en/docs.html).
"Application Program Interfaces" (A.P.I.s) can be defined as a set of tools that enables the communication between different softwares. In this case, Twitter lets us "download" data through an application installed in the server of the WUO, namely "Twitter Capture and Analysis Toolset (DMI-TCAT)" (E. Borra and B. Rieder, 2014). We use this software because, besides being an open source software, it is able not only to extract the Tweet itself in real time: it also attaches rich information to it such as date, location, user, user's bio and other information.

2.2 Data pre-processing

The next phase in the analysis is to pre-process data. This is an important, technical phase in which we clean the data and preprocess it in order to obtain a dataset that can be used for the analysis. To clean the initial dataset of Twitter text we use regular expressions to:
- extract tagged terms ("@[term]", "#[term]", "$[term]"), emoticons and URLs;
- remove tag characters (@, $, #), emoticons and URLs;
- eliminate all non ASCII characters;
- remove trailing empty spaces;
- extract retweets and remove the retweet structure from the text ("RT @[username]:").

Moreover, we keep only those tweets which are written by a user that declared in his user description or in his location to live in the United Kingdom, using one among the following expressions:

"United Kingdom" OR *"UK"* OR *"U.K."* OR *"Great Britain"* OR *"GB"* OR *"G.B."* OR *"England"* OR *"Wales"* OR *"Scotland"* OR *"Northern Ireland"*

In addition to this, we filter Tweets by their provenance, to perform cross-territory analysis. In order to arrive at this territorial division, we start from "Location" column, which is extracted from the User's Location that is in their user profile's Bio. This is the raw data coming from Twitter API, but since the dataset does not provide coordinates to then plot any graph, we use "ggmap" and "raster" packages from R to extract more geographical information from each Tweet, starting from the "Location" cell declared. In particular, this procedure is possible only through a Google Cloud Platform account we create, that queries Google Maps API to extract information (latitude and longitude) from each Tweet's User location (where provided). We then proceed to divide dataset in two other macro-categories: UK constituent countries and UK council areas. Let's look at how data is divided in absolute numbers, by UK constituent country and Countryside versus Urban agglomerates. Tweets with no location are omitted when dividing the dataset.

- *Total (UK): 111251*
- *Divided by region:*
    - *England: 51374*
    - *Scotland: 37338*
    - *Wales: 7290*
    - *Northern Ireland: 2748*
    - *TOTAL: 98750*
- *Countryside/Urban Agglomerates:*
    - *Countryside: 91591*
    - *Urban Agglomerates: 19660*
    - *TOTAL: 111251*

2.3 Taxonomies

Taxonomies are tools which we use in this thesis to better understand the structure of the dataset, giving unique points of view of the data. A taxonomy is a hierarchically ordered list of words that are coming from the same domain of knowledge. At the higher level of this hierarchically ordered list of words we find the most general concept, say "economic variables". To study our subject in a more detailed way, this general concept is further categorized, namely "macroeconomic variables" and "microeconomic variables", and further on because the number of levels depends on the depth of the analysis.

This instrument is very useful for multiple reasons:

- it helps give a detailed representation of a knowledge domain;
- it is shareable among other researchers and professionists;
- taxonomies are hierarchically ordered so that it is possible to do analysis on multiple levels.

For this thesis, we created 7 taxonomies related to economic variables, UK MPs (Members of Parliament), Economic Industries, 100 Greatest Britons of All-Time, Debritt's 500 Most Influential People in UK, Top 50 UK Companies by revenue, and the 25 most influential UK Politicians. We analysed thoroughly the dataset with this tool. These taxonomies will be released publicly for further analysis as open source tools in the "Web Observatory of Uncertainty" website. We hope that other researchers, students as well as companies will enrich these taxonomies or add other ones in order to create an API that is useful for further research. This is the beauty of working with taxonomies: they can be flexibly used with other datasets. Furthermore, by enhancing its use, eventually all knowledge will be classified with this kind of word network.

In figure 3 is the "Economic Variables" Taxonomy that we use for our analysis, with a "Diagonal Network" representation. It shows how we represent the knowledge tree of economic variables.

Figure 3: Diagonal Network of Economic Variables Taxonomy

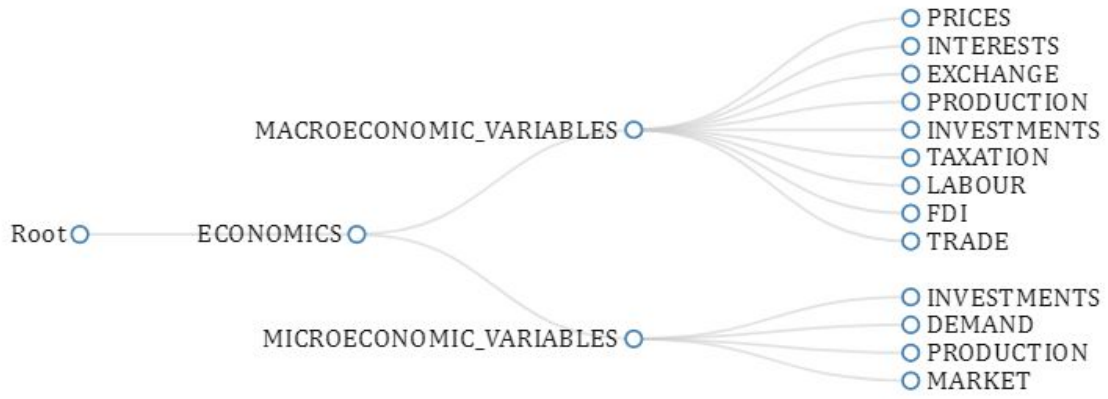


Table 1 is instead a representation of the Quanteda dictionary (in .CAT format) of the taxonomy seen in figure 3. As we can see, the entries at the second-lower level of the hierarchy (for example "INFLATION"), which are called leafs and displayed on the right side of the table below, are associated a set of Regex, each of which has a weight equal to 1. For every leaf of the taxonomy there is one or more regular expression(s) (for example: the leaf called "INFLATION" has three Regex: "RISING_PRICES"). All Regex are case insensitive. These Regex are like logical conditions, based on character matching patterns and used to verify if an observation refers to a specific variable in the taxonomy. This means that, for every observation (tweet) and for every leaf in the taxonomy the script automatically verifies if conditions are met or not. Every time one of those Regex is matched, the value 1 (equivalent to TRUE) will be associated to that leaf of the taxonomy, else the value 0 (FALSE) will be associated to it. Higher levels of the taxonomies act as OR logical operators on the lower level Regex matching conditions. In order to construct this taxonomy, we followed "Ontology Development 101: A Guide to Creating Your First Ontology" by Natalya F. Noy and Deborah L. McGuinness of Stanford University. In order to examine all taxonomies used in the thesis, we suggest to look at the Appendix.

Table 1: Economic Variables Taxonomy

```
ECONOMICS
    MACROECONOMIC_VARIABLES
        PRICES
            INFLATION
                INFLATION (1)
                RISING_PRICES (1)
                PRICE_LEVELS (1)
        INTERESTS
            INTEREST_RATE
                INTEREST_RATE[S]? (1)
        EXCHANGE
            EXCHANGE_RATE
                EXCHANGE_RATE[S]? (1)
                NOMINAL_EXCHANGE_RATE[S]? (1)
                REAL_EXCHANGE_RATE[S]? (1)
        PRODUCTION
            DOMESTIC_PRODUCT
                GDP (1)
                GROSS_DOMESTIC_PRODUCT (1)
                INDUSTRIAL_PRODUCTION (1)
        INVESTMENTS
            PUBLIC_EXPENDITURE
                PUBLIC_EXPENDITURE (1)
                GOVERNMENT_SPENDING (1)
        TAXATION
            TAX
                TAXATION (1)
                TAX[ES]? (1)
                TARIFF[S]? (1)
        LABOUR
            UNEMPLOYMENT
                UNEMPLOYMENT (1)
                UNEMPLOYED (1)
                NO EMPLOYMENT (1)
        FDI
            FOREIGN_DIRECT_INVESTMENT
                FDI (1)
                FOREIGN_INVESTMENT (1)
        TRADE
            TRADE
                TRADE[S]? (1)
    MICROECONOMIC_VARIABLES
        INVESTMENTS
            INDIVIDUAL_EXPENDITURE_AND_INVESTMENTS
                INDIVIDUAL_EXPENDITURE (1)
```

```
        |       |       |       INDIVIDUAL_INVESTMENT[S]? (1)
        |   DEMAND
        |       |   DEMAND
        |       |       |       DEMAND (1)
        |       |       |       QUANTITY_DEMANDED (1)
        |   PRODUCTION
        |       |   PRODUCTION
        |       |       |       PRODUCTION (1)
        |       |       |       QUANTITY_PRODUCED (1)
        |       |       |       OUTPUT (1)
        |       |   WAGE
        |       |       |       WAGE[S]? (1)
        |       |   INPUT
        |       |       |       INPUT (1)
        |       |       |       COST_OF_INPUT[S]? (1)
        |   MARKET
        |       |   MARKET
        |       |       |       CONSUMPTION (1)
        |       |       |       QUANTITY_CONSUMED (1)
        |       |       |       MARKET_SHARE (1)
```

Taxonomies are useful for the majority of analysis that we are going to do. It is only thanks to them that we can analyse the associations to economic variables of interest in the Tweets' dataset. For example, in this thesis we conduct subanalysis using the same dictionary in order to study how certain parts of the dataset (corresponding to different geographic areas of the UK) link uncertainty to specific economic variables showed in the taxonomy.

However, while some taxonomies are generic, such as the one seen above, and can be thus used with a variety of datasets, others are not. For example, we can not use the list of MPs with any type of data set: we created it to specifically study the link among uncertainty, Brexit and leading exponents of the British political world.

Moreover, one can integrate this instrument with other types of data analysis such as semantic networks and co-occurrence network analysis. This helps us understand if and how variables forming taxonomies are related to uncertainty according to twitter's users.

2.4 Analysis

In this section we are going to illustrate the details of our analysis.
This thesis is divided into four main areas:

- frequency analysis;
- taxonomy analysis;
- lexical diversity;
- sentiment analysis.

It is important to better specify the way in which results are presented. In an effort to be more coherent and understandable for the final reader, and to show just the most important results coming from the analysis, we reorganize all results in:

- UK;
- UK's Constituent Members (Scotland, Northern Ireland, Wales, England);
- Countryside vs Urban Agglomerates.

Areas that are considered as part of "Urban Agglomerates": "Cardiff", "Greater London", "Manchester", "Lancashire", "Yorkshire", "Birmingham", "Edinburgh", "Glasgow". All other areas are considered as "Countryside". All the remaining areas are considered as "Countryside".

Starting from the initial dataset composed of 111,251 rows and several variables, we "tokenize" the column containing the "clean tweets". "Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. [...] The tokens become the input for another process like parsing and text mining." (Techopedia.com, 2019). In this way we are able to perform all these analyses, because Quanteda lets us work easily with "stored" sets of tokens and operate with them. We analyse the frequency of these tokens, that is calculating which are the tokens that appear most. We can also select some properties of the tokens, for example select only tokens beginning with "@". This means that our software will measure frequency of the most quoted users, and so on. Through taxonomy analysis, we

exploit the same property of tokenizing the Tweets, but this time we present some "filters" to look for, that are the "word" conditions inside the dictionaries. We are then able to calculate the co-occurrence of the words that make up the taxonomies, and plot a semantic network, accompanied by some calculations that allow us to perform a network analysis. The same property of tokens is helpful to perform Lexical Diversity Analysis, with TTR measure (Type-Token Ratio) and Sentiment Analysis, in which we measure if each token has "positive" or "negative" sentiment, and we even categorise it through emotions. Let's now go a little bit deeper and explain why these four analysis will help us in our final aim.

Firstly, we want to analyse the dataset from a broad perspective, in order to get to know the content from a general point of view. This is the reason why we start from a simple Frequency Analysis, that helps us understand which are the most used words, hashtags, or most quoted users. The beauty of this analysis is that we can make further subsets of data, so we can then compare and show differences on what the most used words, hashtags or quoted users are, in different parts of the UK. In fact, as Ted Kwartler states in his work "Text Mining in Practice with R": "Sometimes merely looking at frequent terms can be an interesting and insightful endeavor. On some occasions frequent terms are expected within a text mining project. However, unusual words or (later in the book as you explore multi-gram tokenization) phrases can yield a previously unknown relationship." (T. Kwartler, 2017). This is exactly what we are doing.

In the next section of the thesis, "Taxonomy Frequency Analysis", we introduce the methodological innovation of Taxonomy Analysis. We are going to analyse the frequency of the words of each dictionary, and put them into a comparison. We shall see which politics, for example, are more quoted and therefore more correlated to uncertainty. Furthermore, through the use of a semantic network, we will put into perspective how social economic variables are linked between each other in this uncertainty scenario.

A semantic network is "a type of data representation incorporating linguistic information that describes concepts or objects and the relationship or dependency between them" (D. Nettleton, 2014). In fact, this graph consist of nodes linked together by lines more or less thick: nodes represent concepts and links connecting the nodes represent the relationships existing among them. Thus, a path in this type of graph is a

sequence of vertices so that from each one there is an edge connected to the next one along the sequence. The underlying assumption views “the meaning of concepts as being determined by their relations to other concepts” (P. W. Foltz, 2001) and this is the reason why we chose to create a Semantic Network.

To analyse our Semantic Network, we measure weighted degree, one of the basic statistical characteristics with which we can understand the Network’s properties. Degree is the total number of connections of a given vertex. In other words, we can define this measure as “ the number of nearest neighbors of a vertex” (J. Borge Holthoefer, A. Arenas, 2010).

We measure it because it shows us how each node of the Semantic Network, and so each concept, is directly influenced by the other surrounding nodes or vice versa.

Moreover we measure modularity which gives us further information about the structure of the graph, in particular it shows us the density of connections linking each node of the network. By dividing a network into modules (groups), high modularity correspond to dense connections among the nodes within modules and scattered connections within others.

One of the few centrality indexes is Betweenness. We measure it because it is important to analyse the stress that each node of the network has to undergo because of its role of intermediary between two other nodes. Betweenness helps us measuring this, by calculating the number of shortest paths between nodes that pass through a given node.

Lastly, we measure Average clustering coefficient of our Semantic Network that is the degree to which adjacent nodes in the network tend to cluster together.

Our next effort is to verify Lexical Diversity inside different sub-groups of the dataset. We do so to understand if the topic of uncertainty is explained with a rich or poor lexicon. Two analysis are performed to get the TTR value (Type-Token Ratio): the first one where we calculate TTR for each Tweet separated, and then take the mean. For the second one we calculate instead the value of TTR for each geographical subdivision, treating each subset as “one big Tweet”. We want to calculate if and in which measure different geographical areas talk about uncertainty.

ⓘ Info Box: TTR calculation and its meaning

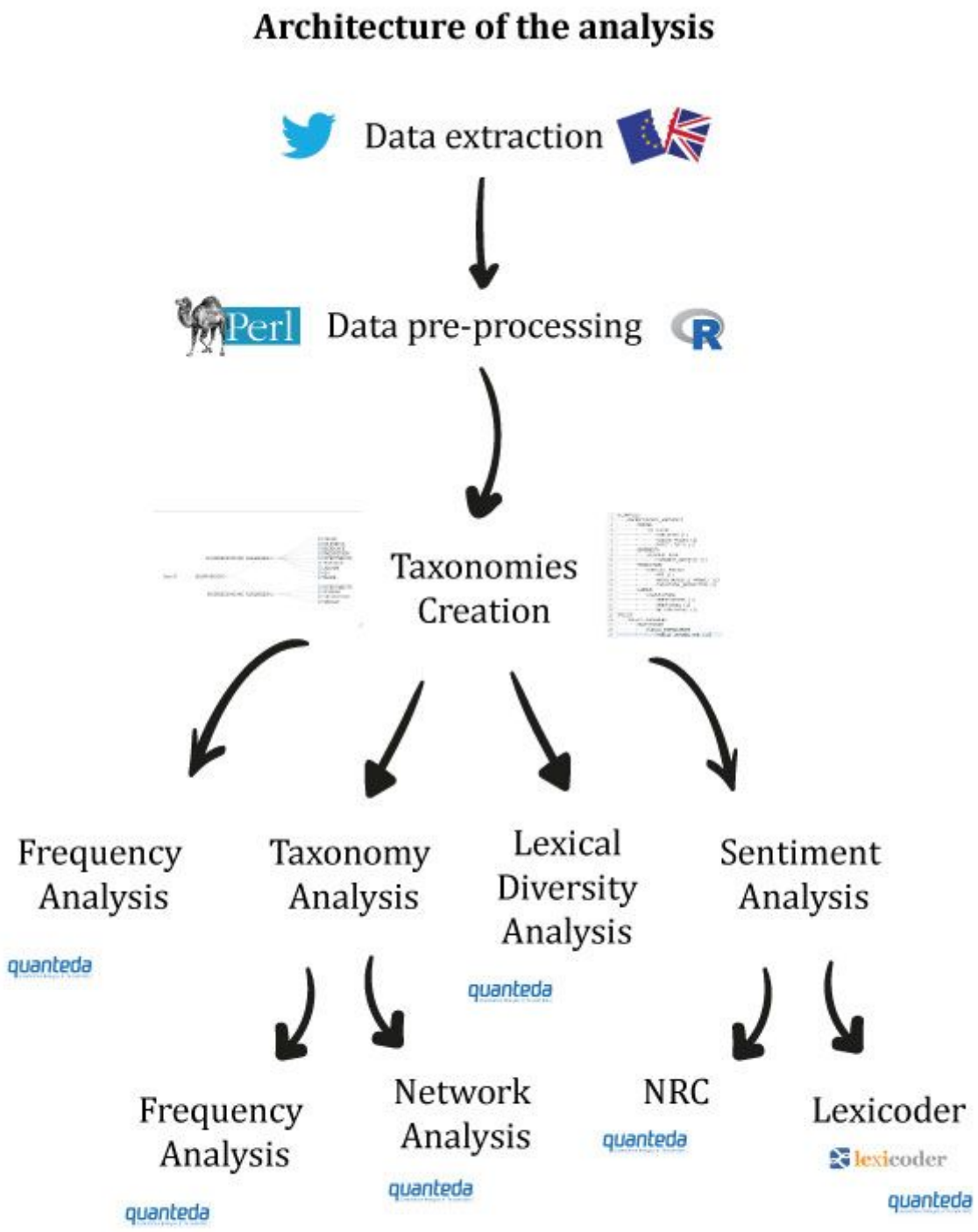
Type-token ratios give a basic insight on the amount of lexical diversity of a corpus of text. It might be a useful (even though very simple) indicator of the complexity of a text. Here "tokens" are the words composing the text/corpus, while "types" are the number of different words within the text/corpus. So for example, in the sentence "I think I will buy a pizza tonight" there are 8 tokens but 7 types, since I repeat the word "I". To calculate TTR we divide the number of types by the number of tokens.

Finally, we perform a twofold Sentiment Analysis. "Sentiment analysis is the process of extracting an author's emotional intent from text." (T. Kwartler, 2017). For the specific purpose of this analysis we select two methods of sentiment analysis: Lexicoder Sentiment Analysis (based on Positive/Negative Analysis of Sentiment) and NRC Sentiment Analysis (based on Emotional Analysis of Sentiment).

Since this part of the work is really technical and we use many libraries for R, we dedicate a final section in the Appendix to illustrate all packages used for each section of the analysis.

Before going to the next section, where we show the results of our analysis, we want to visually illustrate the core steps of the analysis. In this way it is easier to understand the whole architecture behind this work.

Figure 4: Architecture of the analysis



3. Results

3.1 Frequency Analysis

Let's have a look at the general frequency analysis through Figure 5.

Figure 5: Most Used Words in Tweets (UK)

In this analysis we filter out words that occurred less than 2000 times, in order to better understand data and leave out noise. A maximum of 150 words are plotted in each word cloud. From the Figure we already see the most used term, which is "brexit"; but something even more important is the fact that we can "visually" split this set in two parts. The first set of most used words is mostly "political": we see words such as "Theresa" (May), "government", "citizens", "referendum", "political", "negotiations". The second set instead is mostly related to economics: "economy", "businesses", "industry", "investment", "market", "jobs". We can assume then that people strongly link economic variables to uncertainty. And in the particular case of this dataset, to Brexit. We are only scraping the surface of our analysis, but we already see that there is a hint in the right

direction. Let's now see what are the most frequent words used in Tweets, dividing by UK Constituent Members.

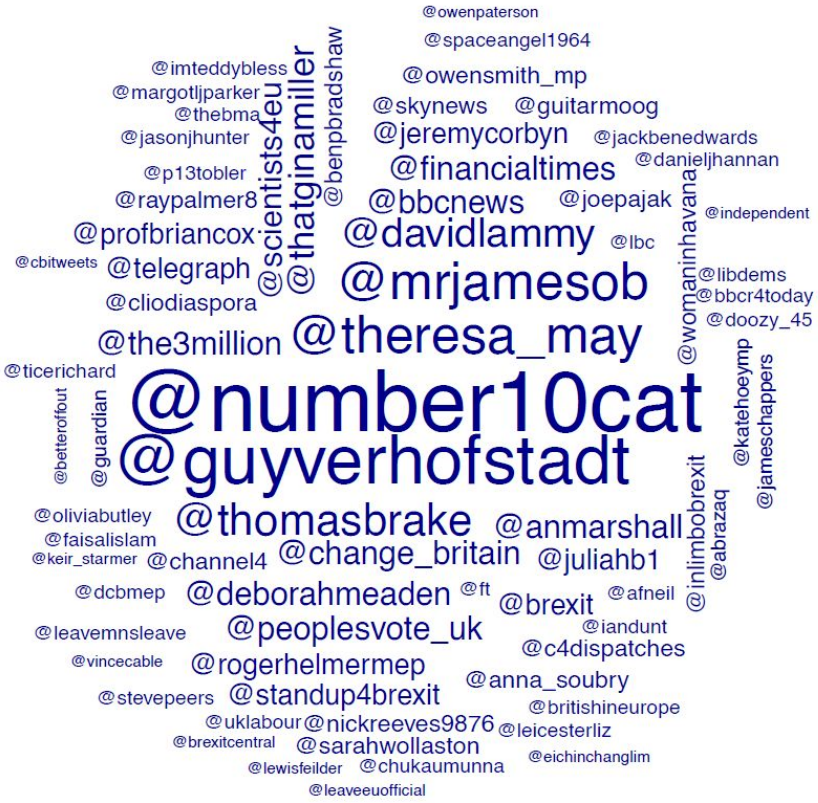
Figure 6: Most Used Words in Tweets (divided by UK constituent countries)

From Figure 6 we understand that the word frequency of the four countries has different peaks. Wales' most frequent words are polarised around "wales", "welsh", "court", "nerves", "businesses" and many words with really negative connotation such as "cancer", "death", "worse". This surely is linked to the "Continuity Bill" welsh and the UK were agreeing on pursuing back in 2018: "The UK Government has confirmed that it will refer the Welsh Government's Brexit law to the highest court in the country" (BBC.com, 17th April, 2018). Northern Ireland's most frequent words are polarised with words like "northern ireland" and "border". The Irish "strong" border is a big debate inside the UK, because of its many implications both in Ireland and Northern Ireland. In the white paper on Brexit published by the UK government, it states that the UK Government's "clearly-stated preference is to retain Northern Ireland's current constitutional position: as part of the UK, but with strong links to Ireland" (HM Government, February 2017). In fact, "Concerns have been raised that the return of a hard border could

jeopardise the Good Friday Agreement. This helped bring the period of violence in Northern Ireland known as "The Troubles" to an end" (BBC.com 9th September 2019). As one may notice, the question of the "Border" is a very felt thing and subject of public debate, as well as a great source of uncertainty for this population. Finally, as we might notice, the plot does not give us clues about any polarization of uncertainty in Scotland and England, since words are proportionally smaller than the other two states: this gives us another stimulus to deepen our analysis.

By filtering all "tokens" beginning with "@" we find the most Quoted Users in Tweets (Figure 7). What stands out from the representation is that the most Quoted User is surprisingly not a politician. Meet "@number10cat", the Chief Mouser to the Cabinet Office. What might seem as a joke, if coming from outside the UK, is in reality a strong "Twitter" sensation, with more than 328,000 Followers as of September, 2019. The cat was adopted by former PM David Cameron in 2011, "and Since receiving the esteemed title of Chief Mouser to the Cabinet Office of United Kingdom of Great Britain and Northern Ireland—the first Downing Street cat to carry the title—he has outlasted Cameron and PM Theresa May, [...], and may have caused a security issue for Donald Trump." (MentalFloss.com, August 2019). He has all around UK media coverage (The Telegraph, The Sun, BBC, SKY and so on) and is mostly followed on Twitter, where he "has an active Twitter parody account, where he comically posts political articles and photos (and has even begun poking fun at his new Downing Street flatmate, Boris Johnson). Sometimes he provides educational information: "England is part of Great Britain (along with Wales and Scotland), which in turn is part of the United Kingdom (along with Northern Ireland)." Other times he just makes cat jokes (see above)" (MentalFloss.com, August 2019).

Figure 7: Most Quoted Users in Tweets (UK)



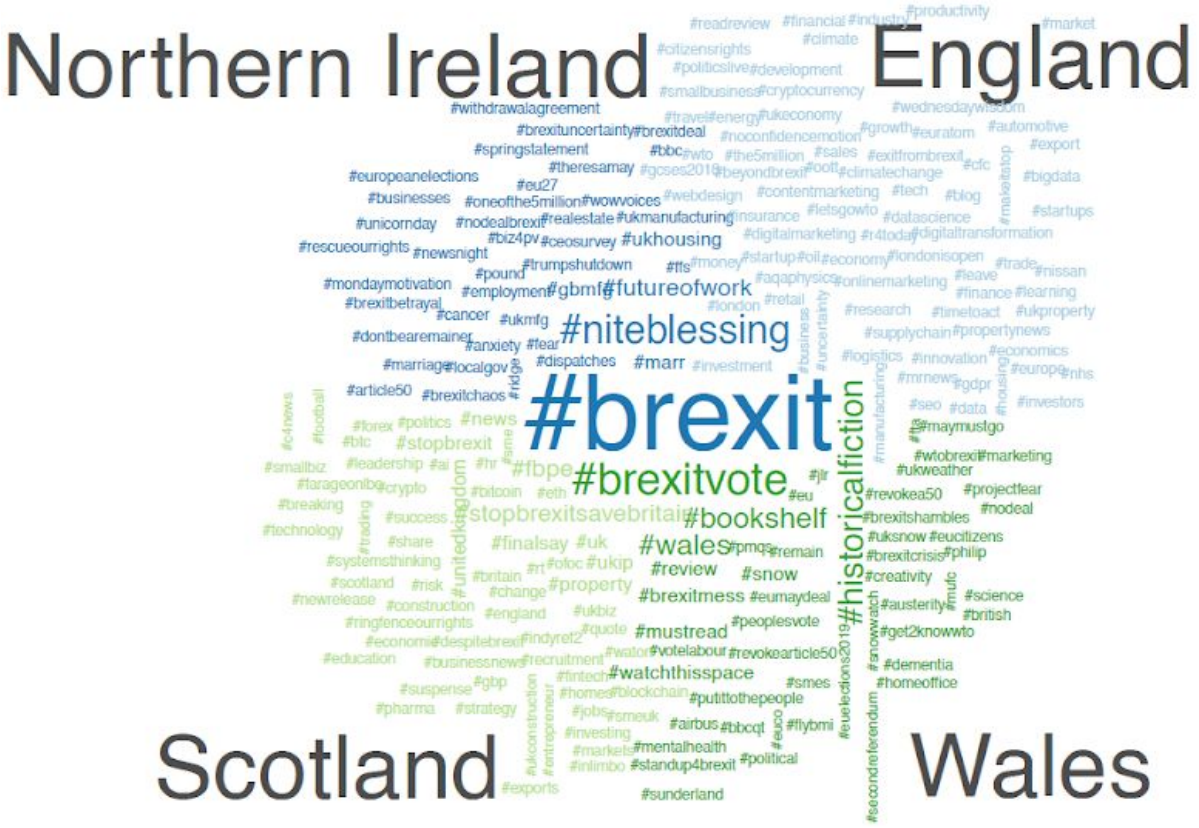
What stands out is that the most “quoted” user on a dataset regarding UK, Brexit and Uncertainty on Twitter is a cat. This shows how British people are when it comes to tough situations, in which they pull out their “British Humour”: “British humour is shaped by the relative stability of British society and carries a strong element of satire aimed at "the absurdity of everyday life". Themes include the class system and sexual taboos; common techniques include puns, innuendo and intellectual jokes.” (L. Laineste, 2014). The second most Quoted User is “@guyverhofstadt”, which is the European Parliament's representative in the Brexit negotiations. Others among Theresa May, Thomas Brake and David Lammy are politicians. James O’Brien (@mrjamesob) instead is a British radio presenter and podcaster.

Figure 8: @number10cat, Larry: Chief Mouser to the Cabinet Office

Figure 9: Most Used Hashtags in Tweets (UK)

Another analysis is done by filtering all “tokens” that begin by “#”: in this way we find the most used hashtags inside our dataset. Here of course, after “#brexit”, we can visually see all the major “hashtag movements” of the period regarding the dataset. An “hashtag movement” happens when an hashtag goes viral because a very high number of users used it in their Tweets, in order to protest or to raise attention towards a specific matter. An example of this phenomenon is “#fbpe”, which stands for “Follow Back Pro EU”, and there’s a very interesting story behind it, that shows how social media nowadays can be an easy vehicule for strong messages, but also subject to manipulations. “The hashtag was first used on Twitter in October by Hendrik Klaassens, a Dutch social media user, : “#ProEU tweeps organize Follow Back Saturdays! Type #FollowBackProEU or #FBPE if you want to get more #ProEU followers. Let’s do this!” in an attempt to build up a network of pro-EU users. [...] With Brexit on the horizon, the idea soon took on a specific twist in the UK, becoming a way for remain voters and pro-EU social media members to identify each other online. Many Brexit supporters have made themselves easy to spot by incorporating flags into their usernames and online biographies, and the aim was to make a similarly easily recognisable signal.” (Theguardian.com, 17th January, 2018). Furthermore, over the next weeks the hashtag has been “hijacked” by many “Brexiteers” who started using the same hashtag but instead with the acronym “Full Brexit Prompt Exit”. What is even more interesting is that this hashtag is now on this wordcloud plot, showing uncertainty at its finest. Other significant hashtags such as “#peoplesvote”, “#putittothepeople”, “#brexitlimbo”, “#brexitmess”, “#revokeA50” or “stopbrexit” let us understand that there have been lots of these “social revolts” during the year, and they tend to be predominantly made by “Remainers”.

Figure 10: Most Used Hashtags in Tweets (divided by UK constituent countries)



When subdividing the analysis of hashtags by UK's constituent countries, we see some steady differences. In fact, hashtags regarding "Remainers", such as "#stopbrexitsavebritain", "#fbpe", "#stopbrexit" are mainly coming from Scotland. From Wales comes the majority of "#revokeA50" as well as "#brexitmess" while from Northern Ireland comes the majority of #brexitbetrayal. This in part reflects how a large part of the population of different members thinks and sees uncertainty in the departure of the United Kingdom from the European Union. These two members in fact voted majority for "Remain" back in 2016 (Scotland 62%, Northern Ireland 55,2%), while England (Remain 46%) and Wales (Remain 47,5%) voted for Brexit.

This Frequency Analysis allows us to have a general overview of the dataset. There are several key takeaways from this section of the analysis. First of all, from a general standpoint, the most frequent words are concentrated among two macro categories, namely "political" and "economical" words. Then, it it interesting to see that Northern Ireland's "border" issue comes up as polarising in the frequency analysis, as well as the

"court" issue in Wales (Continuity Bill). Further filtering the dataset with the most quoted usernames we find "Larry the Cat", an outstanding example of British Humour. It is even more interesting because a great number of mentions comes from media, other from politicians. Finally, hashtags word cloud shows several "Remainers" campaigns to advocate a step back from Brexit. By further subdividing most frequent hashtags by constituent member, we see that a great share of these "Remainers" campaigns comes from Scotland, Wales and Northern Ireland.

3.2 Taxonomy Analysis

3.2.1 Taxonomy Frequency Analysis

Now that we have a general overview of the main words used in the dataset, let's deepen our analysis with the help of taxonomies. Figure 11 gives us a graphical representation of the frequency of Debrett's 500 most influential britons inside our dataset. Level 02 in the taxonomy lets us see only the main category that these Britons are part of, because it is more informative for our research purpose. Unsurprisingly, politicians are the most quoted and debated actors of this climate of uncertainty during the Brexit period. There is no predominant "briton" that is quoted, but taken together, the dominant category is that of politicians.

Figure 11: taxonomy frequency of Debrett's 500 most influential Britons divided by category (Level 02)

Figure 12 depicts the taxonomy frequency of economic variables. Here we choose to look at Level 02 of the taxonomy, since it is the most informative one. The word cloud shows that "trade", "taxation", "production" and "demand" are the most used economic variables. It means that these variables are the most frequently associated with uncertainty. Hence, the main topics of uncertainty, according to our dataset, from May 2018 to May 2019, are the ones we see below.

Figure 12: taxonomy frequency of Economic Variables (Level 02)

“Trade” is the most frequent economic variable, and is indeed among the most known sources of uncertainty. In fact, “The European Union (EU) has about 40 free trade deals, covering more than 70 countries. That means the UK, as a member of the EU, can currently trade with countries like Canada without having to pay taxes on imports (tariffs) on most goods. In the event of a no-deal Brexit, the UK would suddenly lose tariff-free access to these markets and it would have to trade under World Trade Organization (WTO) rules” (BBC.com, 11th September 2019). In the course of all 2019 UK has been signing “continuity” deals privately to ensure a continuity of free trade with other 38 countries outside EU. It is really a source of uncertainty, since UK must ensure continuity in trade; otherwise it could lead to tariffs that could potentially harm its economy, even for Britons: “A 0.2% contraction between April and June is first fall in GDP in six and a half years” (Theguardian.com, 09th August, 2019). This source of uncertainty carries other consequences, such as “taxation”, caused by trade problems, following reduction in tax revenue, and “production”. A study from the Institute for Government of UK carried out in the end of 2018 by G. Tetlow and A. Stojanovic is the theoretical proof of uncertainty in the economic arena. They created this document to sum up what 14 main studies said about Brexit. “The answers range from a prediction that Brexit will boost future economic output by up to 7% through to a prediction that it will reduce it by 18%, compared to what would happen if the UK remained a member of the bloc.” (G. Tetlow and A. Stojanovic, 2018). Highest uncertainty among these 14 official studies seem to be concentrated around 5 areas: trade barriers, FDI, migration, regulations, productivity. Two of these are also the most quoted on Twitter.
Figure 13 represents the same word clouds divided by UK constituent countries, as well as Countryside versus Urban Agglomerates, confirms this trend: the most quoted economic variables are “trade”, “taxation”, “production” and “demand”. There seem to be a slight difference in the terms “production” and “demand”. In the first case (“production”) Northern Ireland seems to emphasize less on it, compared to the other three countries. “demand” instead seems to be stronger in England and Scotland. It is strange how some other economic phenomena such as FDI, which are at the center of the economic debate, aren’t that much quoted in these Tweets.

Figure 13: taxonomy frequency (Economic Variables - Level 02 - divided by Area)

Northern Ireland

Wales

England

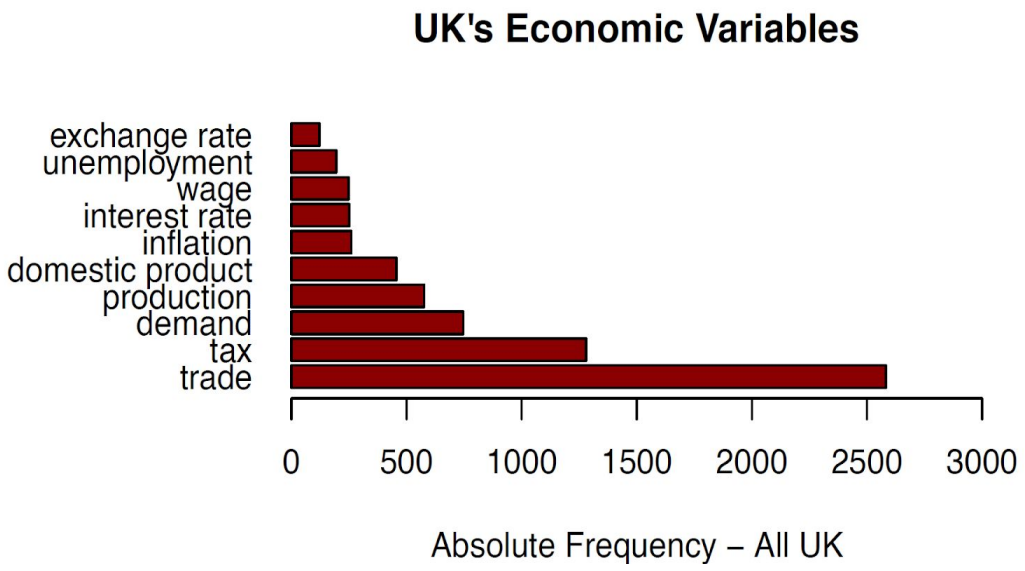
Scotland

Cities

Countryside

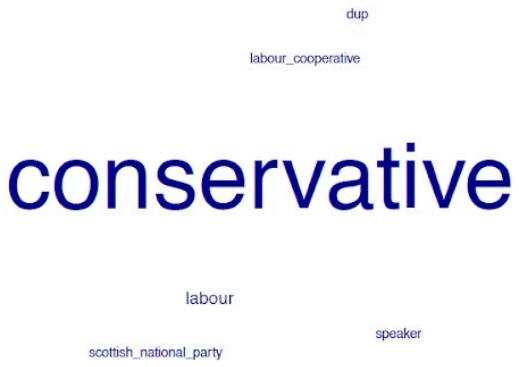
In figure 14 we see a barplot with the absolute frequency of this taxonomy's most used economic variables.

Figure 14: UK's Most Quoted Economic Variables



Another interesting taxonomy is the one that represents all Members of Parliament, divided by Political Party. It is here that we exploit another important feature of taxonomies: the analysis on more levels. Figure 15 depicts "Level 01" of the analysis, where we clearly see that the most "Quoted" MPs are the ones coming from the Conservative party. As we can expect, the most discussed party, and linked to uncertainty, is the same party that supported Brexit.

Figure 15: Taxonomy Frequency of Members of Parliament
(Level 01 - grouped by Political Parties)



Moreover, figure 16 depicts the inferior level of the taxonomy, with the list of all MPs, name by name. The absolute protagonist is Theresa May, from Conservative Party, who has been the political protagonist of the negotiations with EU, up to her stepping down in favour of Boris Johnson on 24th, July 2019. The majority of the other most quoted users comes from the same party: Michael Gove, Philip Hammond, Jeremy Corbyn etc.

Figure 16: Taxonomy Frequency of Members of Parliament (Level 02)

For the next word cloud, in figure 17, we use a taxonomy that groups the top 50 UK companies by revenue. It is really interesting to see if and in which measure they are linked to uncertainty, in the eyes of the public opinion. We see that SSE (Scottish and Southern Energy plc), Reckitt Benckiser Group, BT (British Telecommunications), British American Tobacco, BP (British Petroleum), and Jaguar Land Rover are the most quoted companies. They all are companies with big chunks of operations and revenue outside the UK, both inside the European Union and outside. Of course, for them the climate is still uncertain, because they have to understand whether it is safe for shareholders to stay in the UK, or disinvest. Many companies, after the Brexit

referendum, decided to shift headquarters, business operations or assets out of the UK to Europe. Even banks have transferred more than US$1trillion out of Britain and asset management and insurance companies transferred US$130 billion out of Britain (New York Times, 1st April, 2019). A report from March 2019 from the independent research institute "New Financial" states that out of 269 companies in the banking or financial sector that recently relocated portions of their businesses, 239 of them were because of Brexit. They moved to Dublin, Luxembourg, Frankfurt, Paris or Amsterdam. (W. Wright, C. Benson & E.F. Hamre, 2019).

In particular, Scottish and Southern Energy plc is at the center of a highly debated controversy about the negotiation of UK-EU energy negotiations. An official press release from the official SSE.com website states "When it comes to energy, collaboration with the EU is Imperative". Inside this communication by the CEO Alistair Phillips-Davis points out that "The rationale for a deep, comprehensive and collaborative energy relationship with the EU is not economic alone, but critical if we are to decarbonise our economy". Furthermore he states: "the UK and EU should continue to collaborate on delivering large, ambitious energy projects for mutual benefit. [...] There is potential to connect and provide power between the UK, Germany, Denmark, Netherlands and Norway. Working with our European colleagues we can make the North Sea low carbon grid a reality. [...] These are complex matters, but the sooner the parties to the negotiations on the future UK-EU relationship on energy are able to agree a way forward, the better it will be for efforts to take forward the next stages in decarbonising our economy." (SSE.com, 02nd February 2018). Evidently the tensions between SSE, UK's broadest-based energy company, and the government are a big cause of uncertainty in the eyes of Twitter's Users.

Reckitt Benckiser, instead, is widely reported as an example of companies benefiting from Brexit. First of all, RB is an internationally-focused business, and as such, has a portfolio of products that is beloved in dozens of countries (10% approx. of sales come from UK). They report in sterling, which depreciated by around 12% since the referendum, so that would be a benefit from a positive currency translation, for financial figures purposes. Export should also increase, giving them more competitivity. This in turn could allow them to drive their revenue even higher over the medium run. The reason why these are the most quoted companies might be explained by a

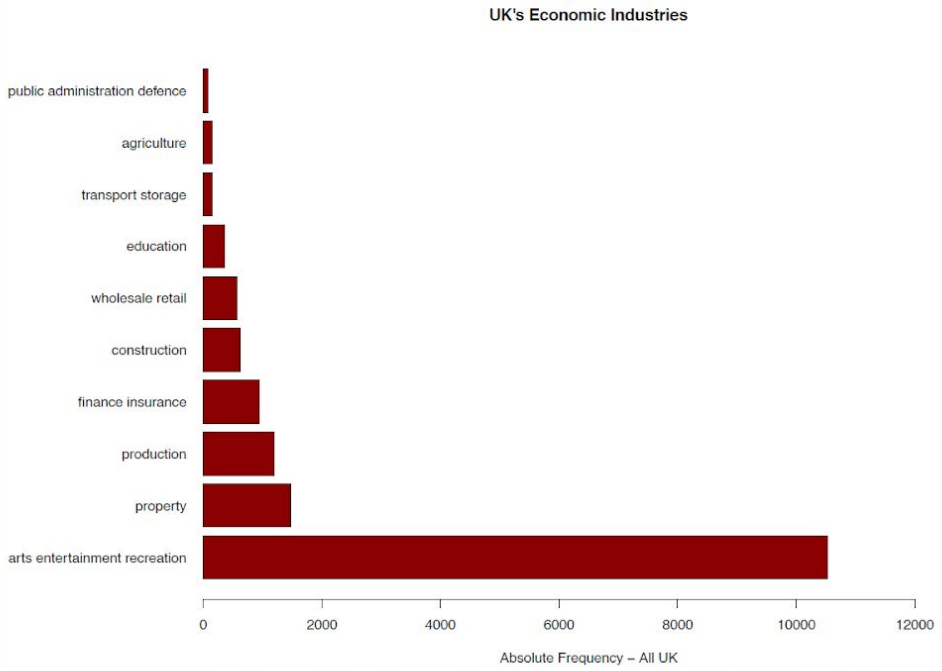
phenomenon called “mere-exposure effect” which was studied in human decision-making: “[...] people invest in the familiar while often ignoring the principles of portfolio theory.” (G. Huberman, 2015). That is, there is a human tendency to invest in certain companies just because they are more familiar with them, even though international markets offer similar or better alternatives. In this case Twitter users act like stock investors. For instance SSE and Reckitt Benckiser Group might not have the worst uncertainty problem out of all 50 companies, but they surely are among the most known companies to the general public. In fact, other companies such as KAZ Minerals and Blackrock funds might also be linked to uncertainty related to Brexit, but the majority of the public is not really familiar with them.

Figure 17: Taxonomy Frequency of Top 50 UK Companies by Revenue (Level 02)

Lastly, the taxonomy regarding “Economic Industries” gives us another hint at uncertainty: as we see in figure 18, the most quoted economic industry related to uncertainty is the entertainment business. In fact, a report made in 2017 by the UK Government stated that UK media industry now contribute £92 bn net per year, about 9% of all services exported from the UK and 1/11 of jobs. This industry is also growing at a rate which is double that of the economy. This economic sector is very likely to be correlated with uncertainty. On Brexit, “UK-based broadcasters and producers will lose the benefit of a number of favourable EU laws. Simply incorporating these laws into

domestic UK legislation will not provide a solution since they require reciprocity from the EU, something the UK cannot control. Perhaps the most critical of these laws is the Audiovisual Media Services Directive (Directive 2010/13/EU) and its successors (the AVMSD). This Directive allows broadcasters to operate across the EU if they satisfy the regulatory requirements, and are licensed, in the Member State in which their services originate (the so-called "Country of Origin" or "COO" principle). Many international - particularly US - broadcasters take advantage of this regime, basing their EU operations in the UK and being licensed by Ofcom. Indeed, Ofcom currently licenses more than half of the 2,200 channels broadcast EU-wide. This has been estimated as a business worth more than £5 billion per year in the UK." (Taylorwessing.com, 2018).

Figure 18: Taxonomy Frequency of Economic Industries

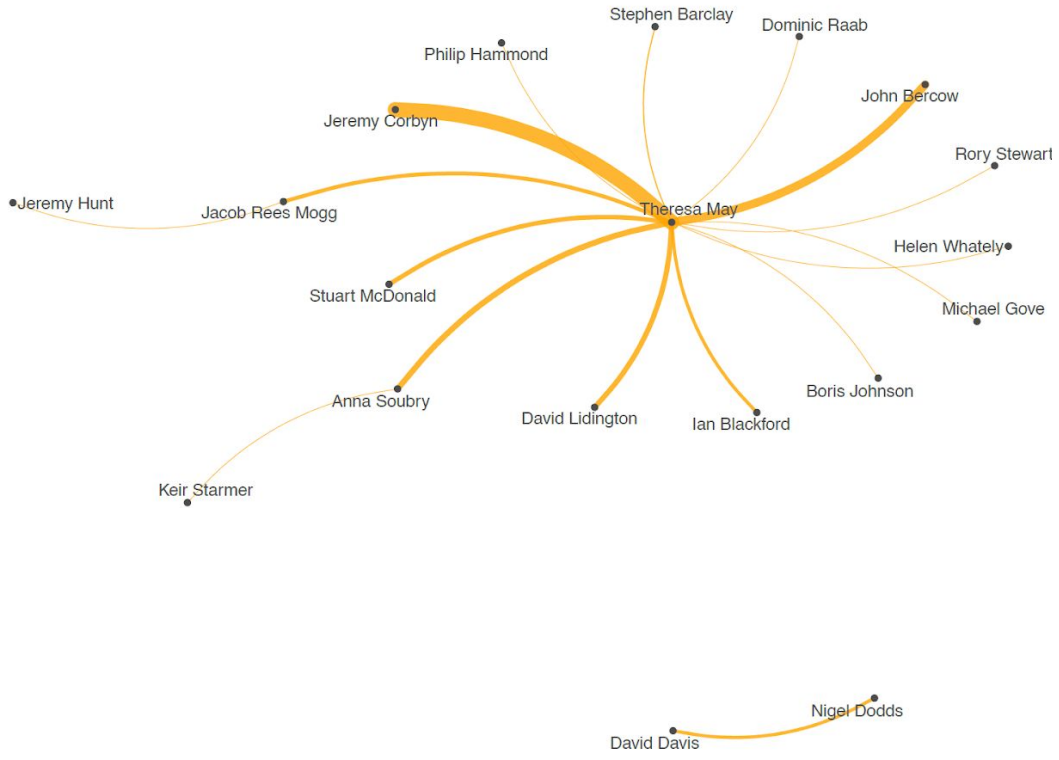


This section enlightens many interesting points. The main one concerns the taxonomy frequency of Economic variables: the most frequent economic variables linked to uncertainty are “trade”, “taxation”, “production” and “demand”. This result is confirmed even when subdividing the data set into the four UK countries and into urban agglomerates-countryside. The taxonomy from Debrett’s 500 most influential Britons shows that politicians are the most quoted Britons, while a further investigation, with the help of the taxonomy of Members of Parliament (Level 01) shows that conservatives are the most quoted Members of Parliament. We can also go deeper than that, with Level 02 of the same taxonomy showing us that Theresa May is by far the most quoted MP. Lastly, it is very interesting to see how people from Twitter associate the Top 50 UK Companies by revenue to Brexit and Uncertainty. The most quoted ones are SSE, Reckitt & Benckiser, BT, BP, British American Tobacco for multiple reasons aforementioned. For example, these are companies that operate all over the world: the majority of them has headquarters in the UK and now they are facing tough decisions like deciding whether or not to relocate outside the UK. It is interesting now to put all these frequencies in perspective, with the help of semantic networks and network analysis.

3.2.2 Taxonomy Network Analysis

We begin our analysis of the semantic network by looking at figure 19: it is the semantic network of Members of Parliament**.** At the center of the Network there is Theresa May, which is related to other important politicians. The strongest links are between Jeremy Corbyn (leader of the opposition, Labourist), and John Bercow (member of the Conservative Party). What stands out is that all politicians tend to be exclusively linked with Theresa May: they are not linked with each other, except for singular cases such as Jeremy Hunt and Jacob Rees Mogg, Anna Soubry and Keir Starmer, but mostly, David Davis and Nigel Dodds.

Figure 19: Semantic Network (2018 UK MPs - Level 02 - UK)



From a more technical point of view (figure 20), both measures of Degree and Betweenness indicate that Theresa May has the strongest position in the network. In

fact the node of Theresa May is the central one in the representation. The average clustering coefficient is zero: in fact, adjacent nodes in the network don't cluster together. This means that people tend to associate Theresa May with many politicians, but other politicians do not correlate between each other as much to create other links between them. This is probably due to the fact that she has been Prime Minister and therefore she is the most known actor in the political scene.

Figure 20: Network Analysis (2018 UK MPs - Level 02 - UK)

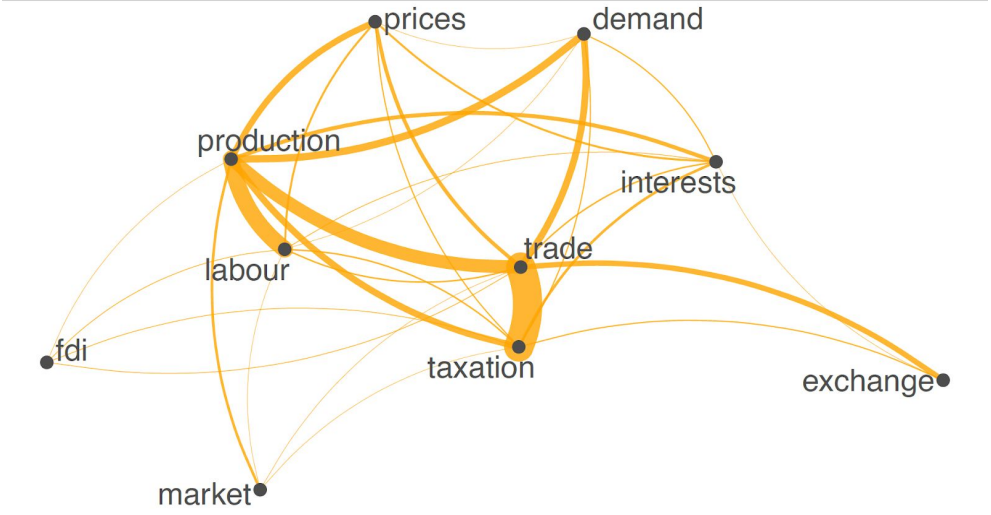
UK MPs	Degree
theresa may	80
jeremy corbyn	24
john bercow	13
anna soubry	10
david lidington	8
jacob rees mogg	7
stuart mcdonald	7
david davis	5
nigel dodds	5
ian blackford	5

UK MPs	Betweenness
theresa may	50
anna soubry	6
stephen barclay	0
david davis	0
michael gove	0
philip hammond	0
jeremy hunt	0
boris johnson	0
david lidington	0
dominic raab	0

Modularity	Avg Clustering Coeff.
0.108336672186852	0

The next semantic network (figure 21) shows us how economic variables are related to each other in the dataset. At a first glance, it seems like trade and taxation have a quite strong connection, like production and labour do. In fact, these variables tend to influence each other also in the real economic world. Strangely, FDI seems to be pretty isolated from the network.

Figure 21: Semantic Network (Economic Variables - Level 02 - UK)



Technically speaking, trade has the highest Degree by a great margin (222): it is the most connected node of the network. This means that this is the most quoted word within the taxonomy along with other words. Then follow production and taxation. FDI, as we were previously stating, is stuck at 9 connections. Betweenness shows that taxation is instead the most stressed node in the network, by a small margin with respect to trade. The most central node in the network instead appears to be prices, followed by interests. Modularity this time is zero, which means that there is a low number of connections among the nodes within modules and dense connections within others.

Figure 22: Network Analysis (Economic Variables - Level 02 - UK)

Economic Variables	Degree
trade	222
production	201
taxation	158
labour	83
demand	58
prices	53
interests	42
exchange	23
fdi	9
market	9

Economic Variables	Betweenness
taxation	3.44188448
trade	3.30230859
production	2.61637931
labour	1.60925520
interests	0.03017241
prices	0.00000000
exchange	0.00000000
fdi	0.00000000
demand	0.00000000
market	0.00000000

Modularity	Avg Clustering Coeff.
0	0.825396825396825

Figure 23: Semantic Network (Economic Variables - Level 02 - divided by Area)

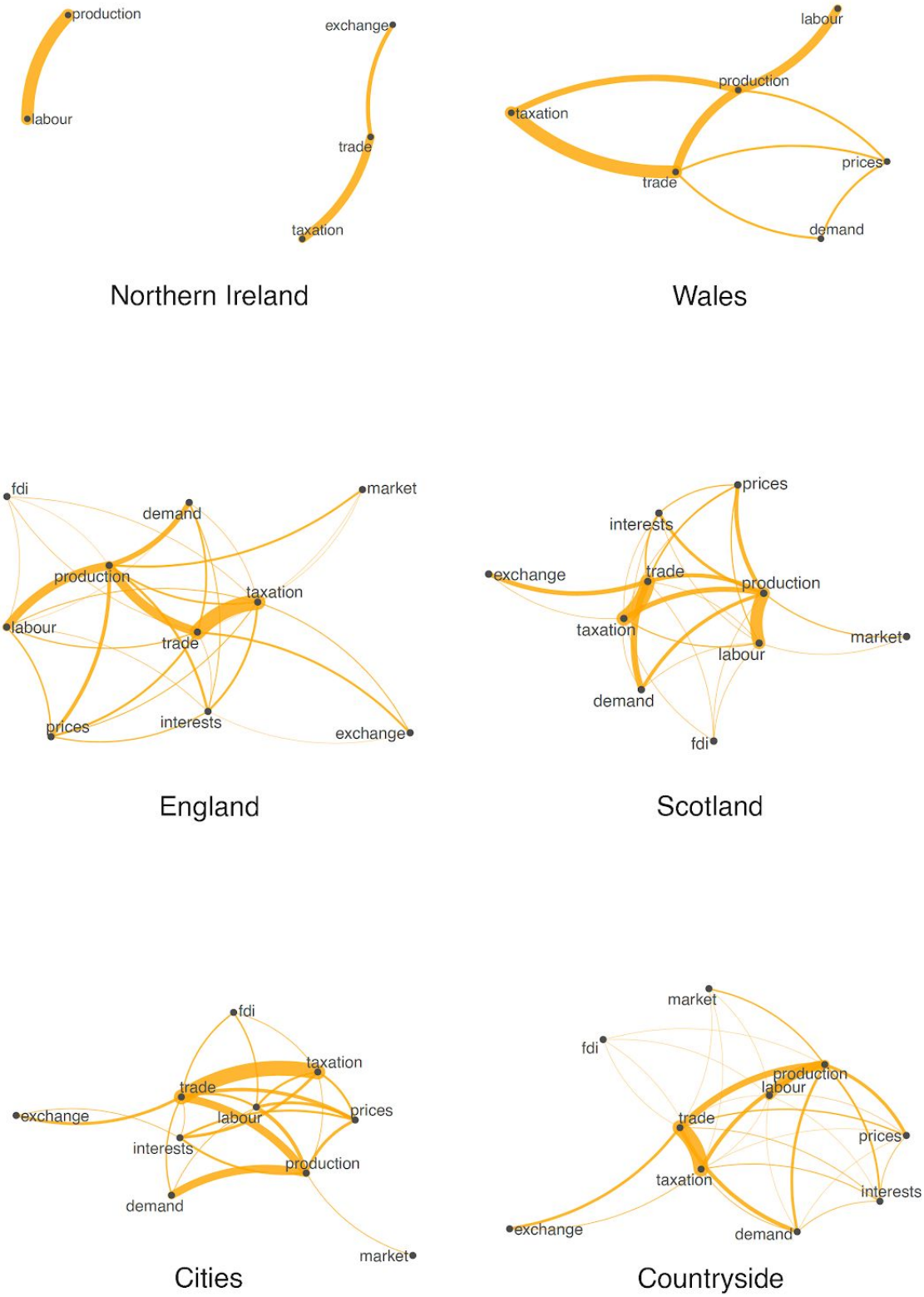
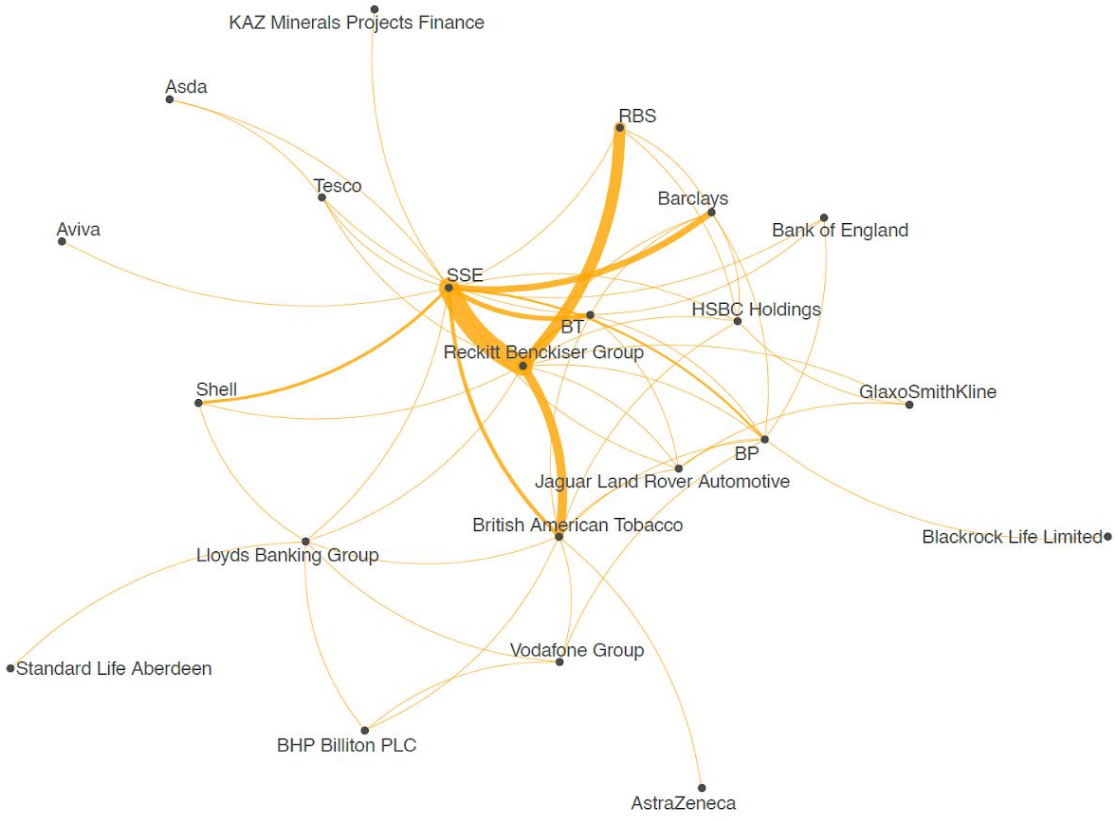


Figure 23 shows 6 semantic networks, all representing the most quoted economic variables from the economic variables taxonomy, but this time divided by the 4 constituent countries and by Urban Agglomerates/Countryside. We choose this kind of subdivision because we want to see differences in perception of uncertainty in different parts of the United Kingdom. We clearly see that "trade", "taxation" and "production" are the strongest links in all areas except from Northern Ireland, where they link "taxation", "trade" and "exchange" in a network which is well separated from "production" and "labour". "demand" has a stronger link with "production" in Cities (or Urban Agglomerates) than it has in the Countryside. Scottish people tend to correlate more "exchange" and "trade" than the other three countries.

The semantic network of UK's Top 50 companies by revenue informs us that SSE, Reckitt Benckiser Group, British American Tobacco and RBS are the strongest nodes in the network.

Figure 24: Semantic Network (Top 50 UK companies by revenue - UK)



In fact, from the network analysis we can clearly state that these companies have the highest level of degree by a large margin. SSE instead is the most stressed node of the network (betweenness = 36,22), while there appears not to be a unique central node in the network. Moreover, a modularity level of 0.008 indicates that there are scattered connections among the nodes. Lastly, average clustering coefficient seems to have a pretty high value, so it means that adjacent nodes in the network tend to cluster together. In other words, people writing about uncertainty in relation to UK companies tend to talk about more companies at the same time or, better, in a unique Tweet.

Figure 25: Network Analysis (Top 50 UK companies by revenue - UK)

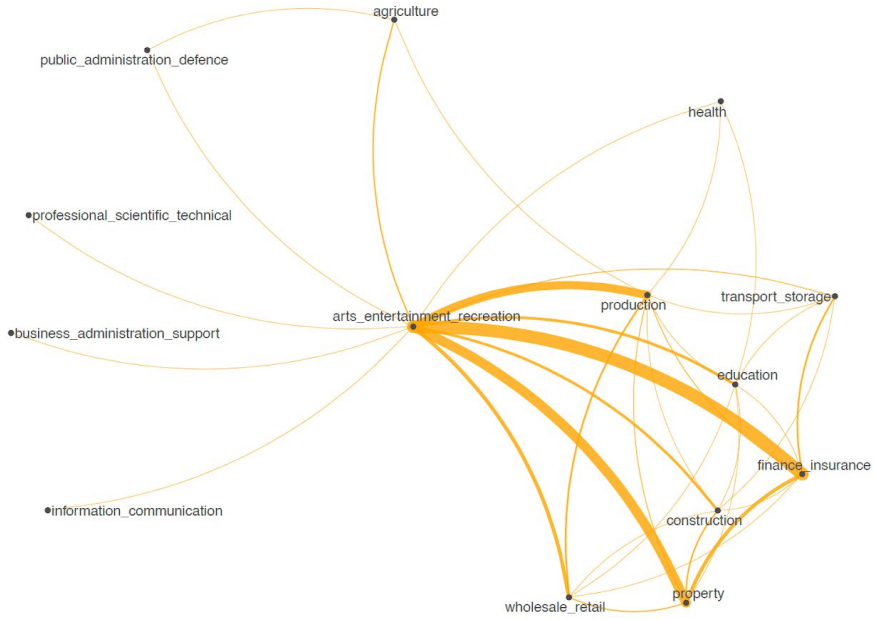
Top 50 UK Companies	Degree
Reckitt Benckiser Group	719
SSE	678
British American Tobacco	223
RBS	200
BT	141
Barclays	112
BP	69
Shell	51
Jaguar Land Rover Automotive	19
Bank of England	14

Top 50 UK Companies	Betweenness
SSE	36.22787506
Reckitt Benckiser Group	10.38190018
British American Tobacco	6.00000000
BT	4.33515789
Lloyds Banking Group	3.00112317
Jaguar Land Rover Automotive	3.00000000
Vodafone Group	1.00000000
GlaxoSmithKline	0.04000000
Barclays	0.01394369
Shell	0.00000000

Modularity	Avg Clustering Coeff.
0.00864050909876823	0.680880230880231

As concerns Economic industries (Figure 26), the biggest links are among Arts, Entertainment and Recreation (Media Industry) respectively with production, finance and insurance, and property. Media industry (arts_entertainment_recreation) seems to have a central role on the network.

Figure 26: Semantic Network(Economic industries - Level 02 - UK)



3.3 Lexical Diversity Analysis

We investigate lexical diversity to discover whether the Tweets were more or less complex from a lexical point of view, and so to have another, quick measure of how people describe uncertainty. We run two analysis trying to get the value of TTR: the first one keeping Tweets separated, so that the value of TTR reflects the lexical diversity of each single Tweet. The second analysis, instead, measures TTR by calculating tokens of Tweets all together (ex. “all tokens coming from England”, “all tokens coming from the countryside”). Results can be found in table 2 and 3.

Table 2: TTR (UK constituent countries)

	England	**Scotland**	**Northern Ireland**	**Wales**
Mean TTR (Tweets separated)	0.969	0.968	0.969	0.973
TTR (One big Tweet)	0.062	0.069	0.196	0.128

Table 3: TTR (UK Urban Agglomerates vs countryside)

	Urban Agglomerates	Countryside
Mean TTR (Tweets separated)	0.968	0.970
TTR (One big Tweet)	0.101	0.047

We clearly see that when calculating TTR Mean value (keeping Tweets separated) the value of TTR is really high. In fact TTR (Type Token Ratio) is calculated as N/R, where N is the total number of tokens, and R is the number of different types. It is really difficult that in the same Tweet a word is highly repeated. A completely different result we have when considering TTR of all tokenized Tweets from each UK's constituent country, or Urban Agglomerates - Countryside. We consider all tokens as "One Big Tweet" coming from the same subset of the dataset. The result, as one should expect, is the opposite, meaning that there is high Lexical Diversity. The results differ more than the previous ones by a fair amount, but this is so because usually "Lexical diversity (LD) measures have been known to be sensitive to the length of the text" (R. Koizumi, 2012). In fact, the longer the text, the more probability that the Lexical Diversity will go down. What we do so is to make a linear regression with a scatterplot on a log log scale of TTR and N. What we are really interested in is the deviation from the line, in this case we see if the diversity is coherent when dataset changes (ex. Tweets coming from Wales are 14% of those coming from England, so one should expect that Lexical Diversity changes proportionally to the size of the dataset). Finally, we see whether deviations lie within the line (that would show that there is no difference in Lexical Diversity) or if they lie above or lower to the regression line. In the next figures, we abbreviate the word "Urban Agglomerates" with "Cities" for simplicity. Figure 27 depicts the log log linear regression with the 4 UK Constituent Countries, while figure 28 just with Urban Agglomerates and Countryside. Finally, figure 29 puts the two precedent analysis together. From the analysis it seems like there is not a very significant outlier, there are just some slight deviations.

Figure 27: Log Log Linear Regression of TTR and N° of Tweets (divided by UK Constituent Countries)

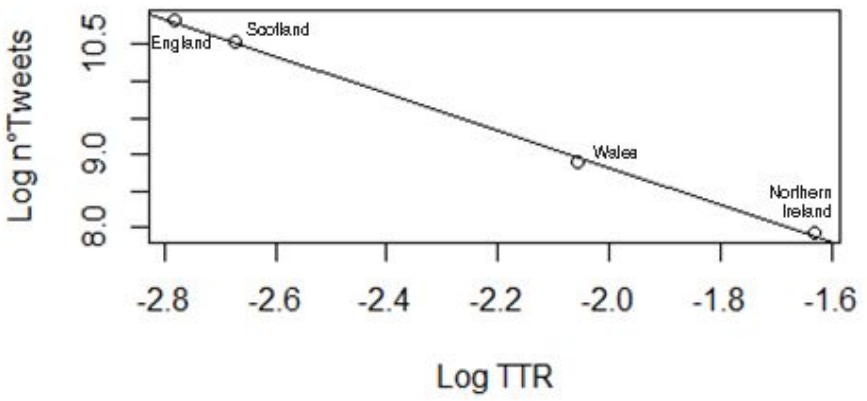


Figure 28: Log Log Linear Regression of TTR and N° of Tweets (divided by Urban Agglomerates and Countryside)

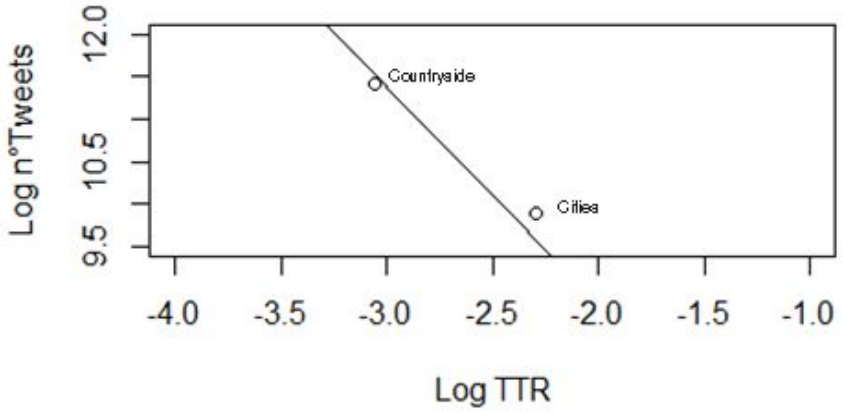
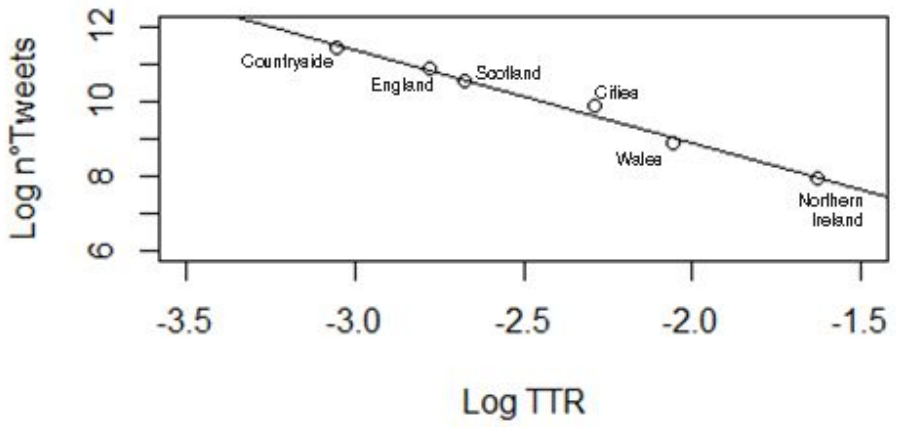


Figure 29: Log Log Linear Regression of TTR and N° of Tweets (plot of Urban Agglomerates, Countryside and UK Constituent Countries)



In fact, we proceed to calculate residual errors. In R, this is done with summary of lm() function. As we see from table 4 the main outliers are Cities (Urban Agglomerates) with 0.273, and Wales with -0.131.

Table 4: Residual Errors from Linear Regression (Figure 29)

	Countryside	England	Scotland	Cities	Wales	Northern Ireland
Residuals	-0.0848	0.024	-0.030	0.273	-0.131	-0.051

Figures below are a graphical representation of TTR Lexical diversity over the timespan of the dataset.

Figure 30: Lexical Diversity Analysis (TTR per Tweet, UK)

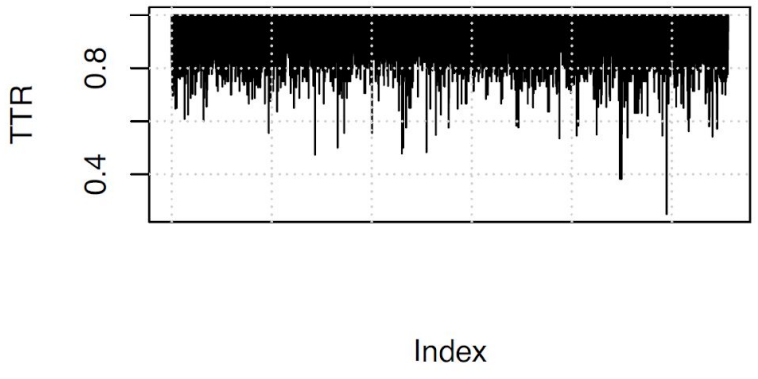


Figure 31: Lexical Diversity Analysis (TTR per Tweet, England)

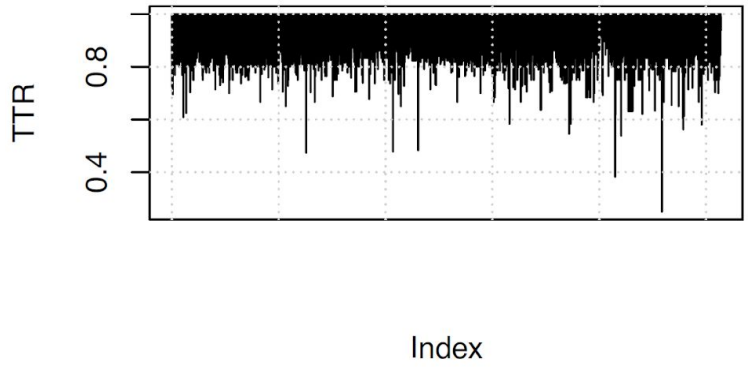


Figure 32: Lexical Diversity Analysis (TTR per Tweet, Northern Ireland)

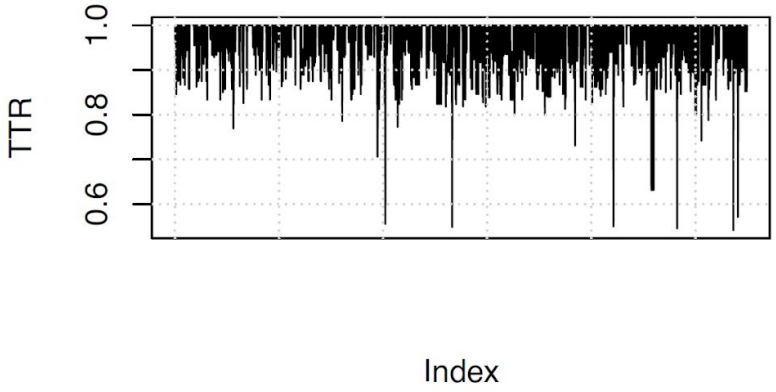


Figure 33: Lexical Diversity Analysis (TTR per Tweet, Scotland)

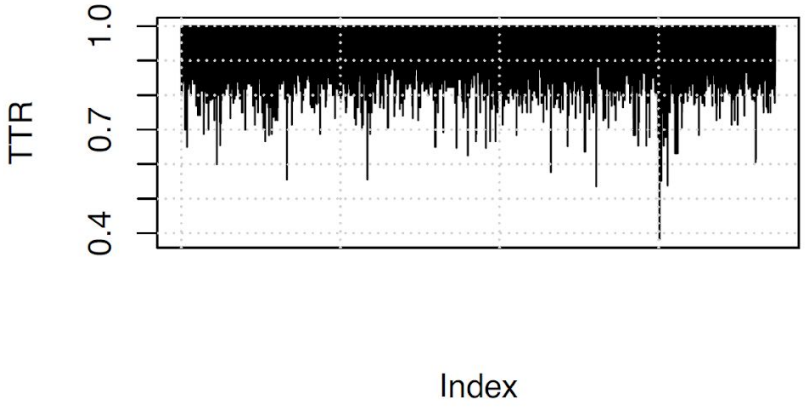


Figure 34: Lexical Diversity Analysis (TTR per Tweet, Wales)

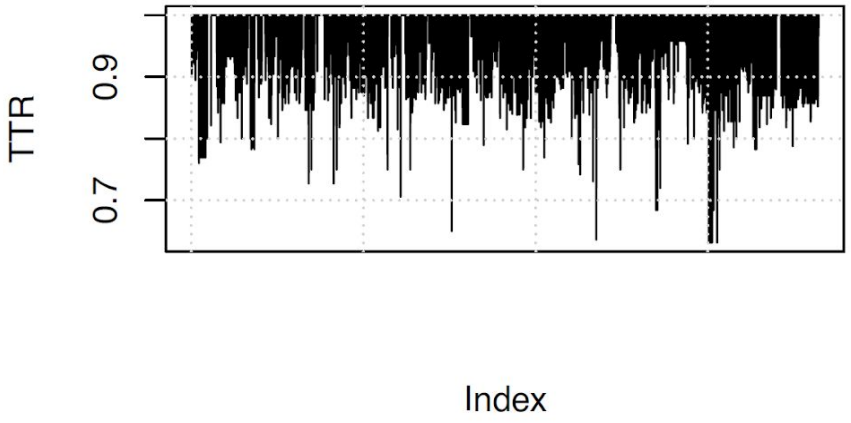


Figure 35: Lexical Diversity Analysis (TTR per Tweet, Urban Agglomerates)

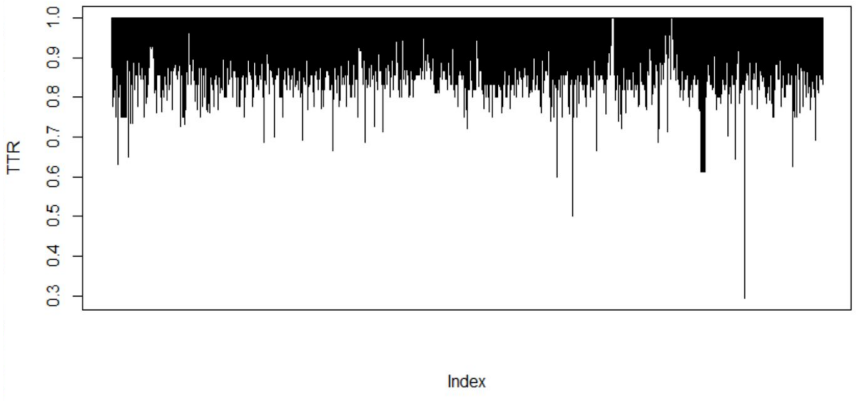
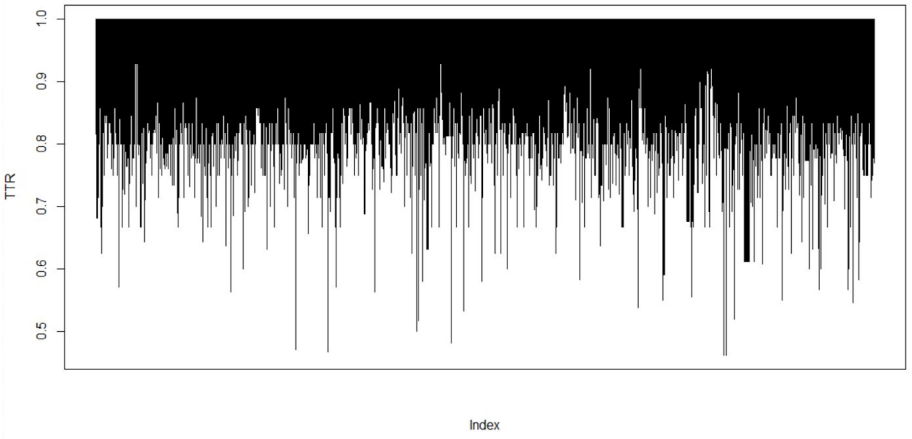


Figure 36: Lexical Diversity Analysis (TTR per Tweet, Countryside)



In this section we have analysed Lexical Diversity, to measure how many different words are used in a text. We clearly see that it seems to be coherent among our subdivisions of the subset (when taking the log log linear regression of number of Tweets and TTR), with the exception being “Urban Agglomerates” and “Wales”. Urban Agglomerates have a higher value of TTR, not in absolute terms (with respect to others) but considering the size of this data subset. Viceversa is true for Wales. This means that Tweets about uncertainty that come from Urban Agglomerates use a more varied lexicon, in relative terms. The Lexicon is less varied for Tweets that come from Wales.

3.4 Sentiment Analysis

3.4.1 Lexicoder Sentiment Analysis

We proceed to analyse the sentiment of these Tweets, firstly with the Lexicoder Sentiment. Figure 37 illustrates the distribution of Sentiment throughout UK. The mean sentiment is -1.27. This means that perception of uncertainty and Brexit is shifted mainly towards negative feelings (mean sentiment: -1.27).

Figure 37: Lexicoder Sentiment Analysis (UK)

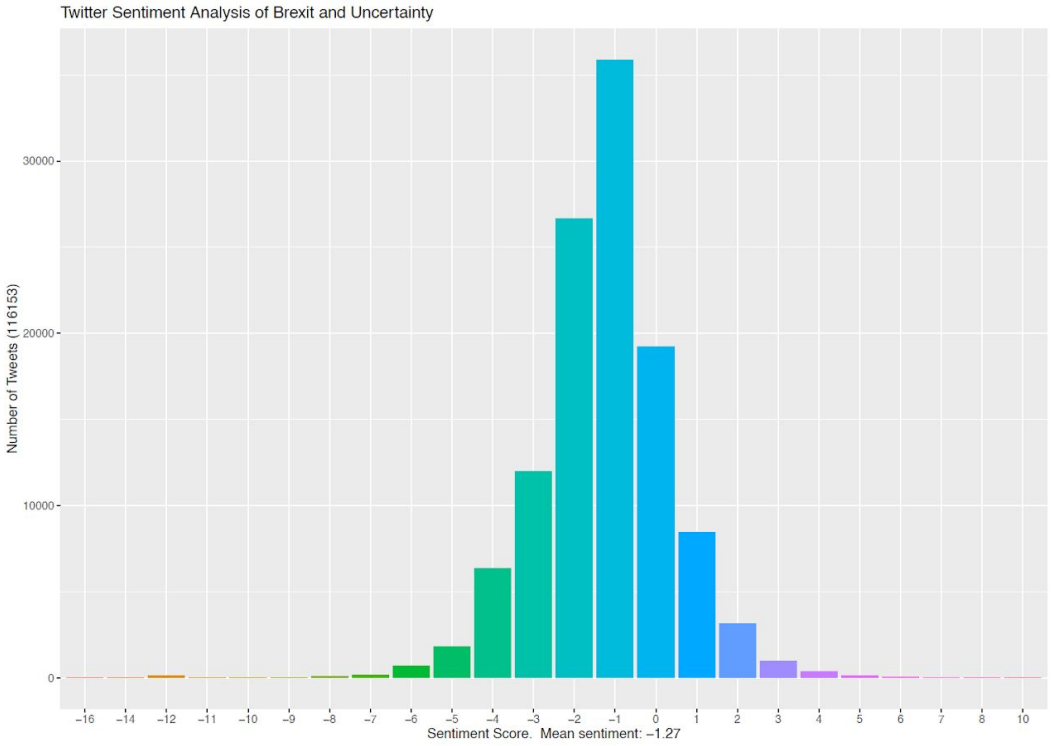
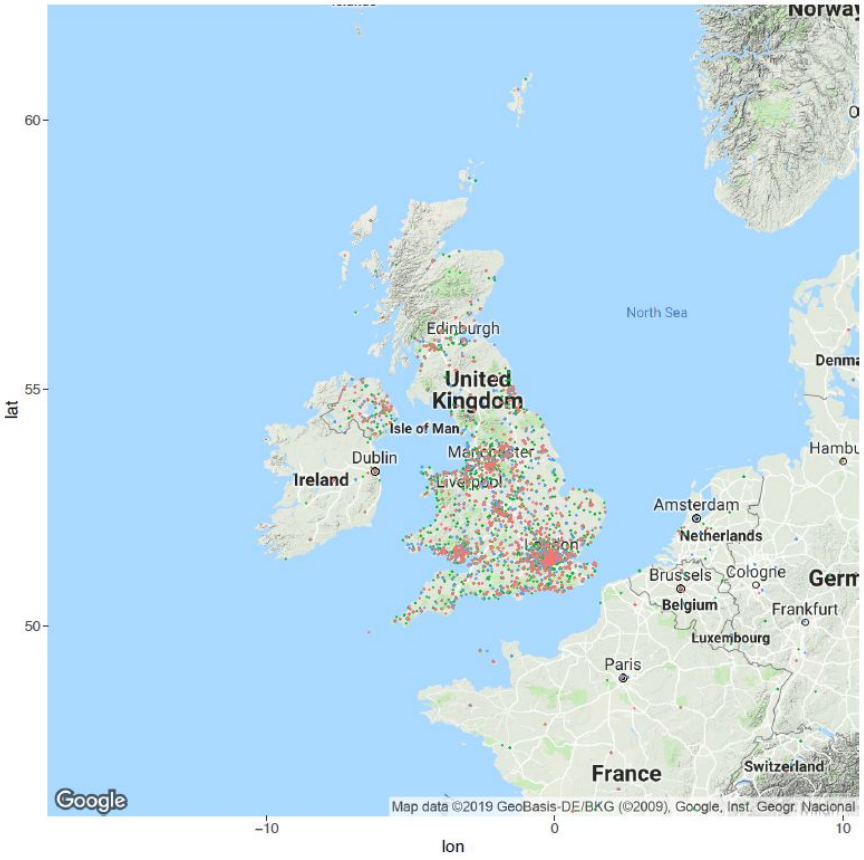


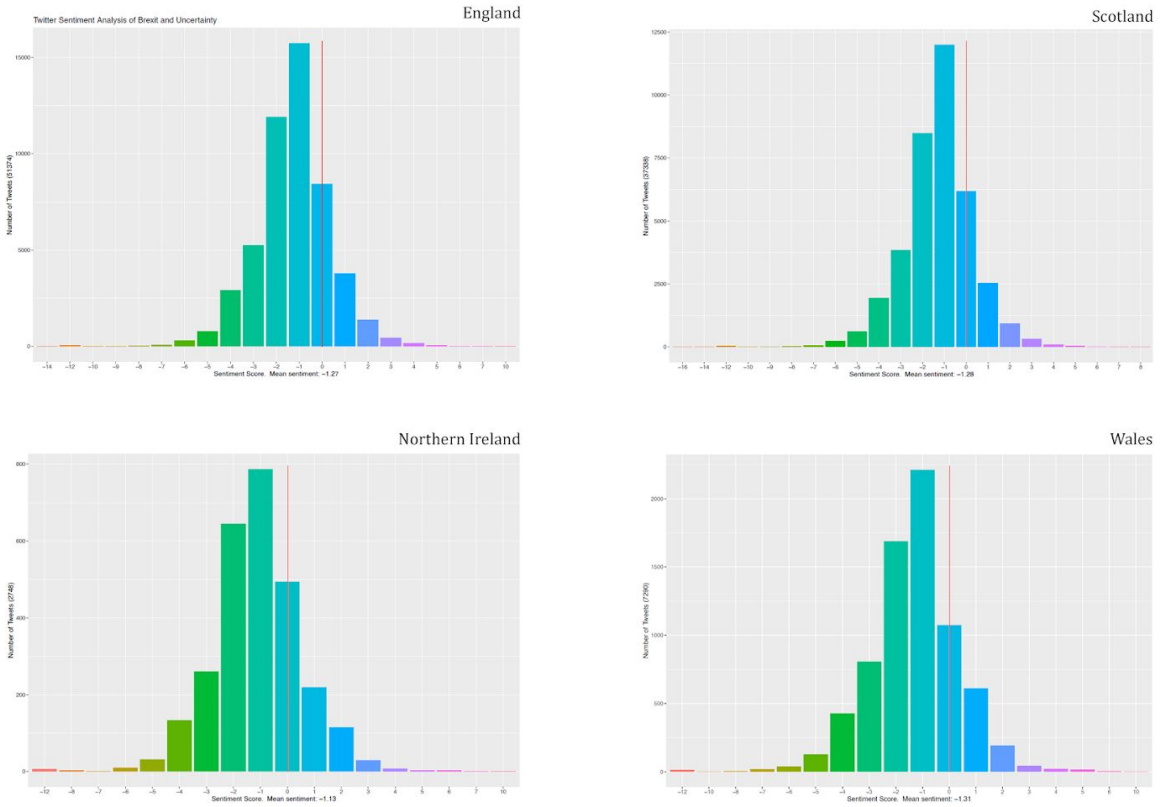
Figure 38: Geographic Location of Lexicoder Sentiment Analysis (UK)



In green, we see positive sentiment. Red shows negative sentiment, while blue shows neutral sentiment. As we clearly see, negative sentiment is mainly concentrated in Urban Agglomerates such as Manchester, Liverpool and London.

If we divide our dataset in the four constituent countries, as in figure 39, we see that the situation is pretty much similar in all four constituent countries. The lowest value of mean sentiment is found in Wales, with -1.31. We can see also graphically two small bars at the very left of the Wales barplot, indicating a small amount of really negative sentiment that could be the cause of this such low value. Next up there's Scotland, with -1.28 mean negative sentiment, next to England (-1.27). The "least" negative is Northern Ireland, with a mean of -1.13.

Figure 39: Lexicoder Sentiment Analysis (divided by UK constituent countries)



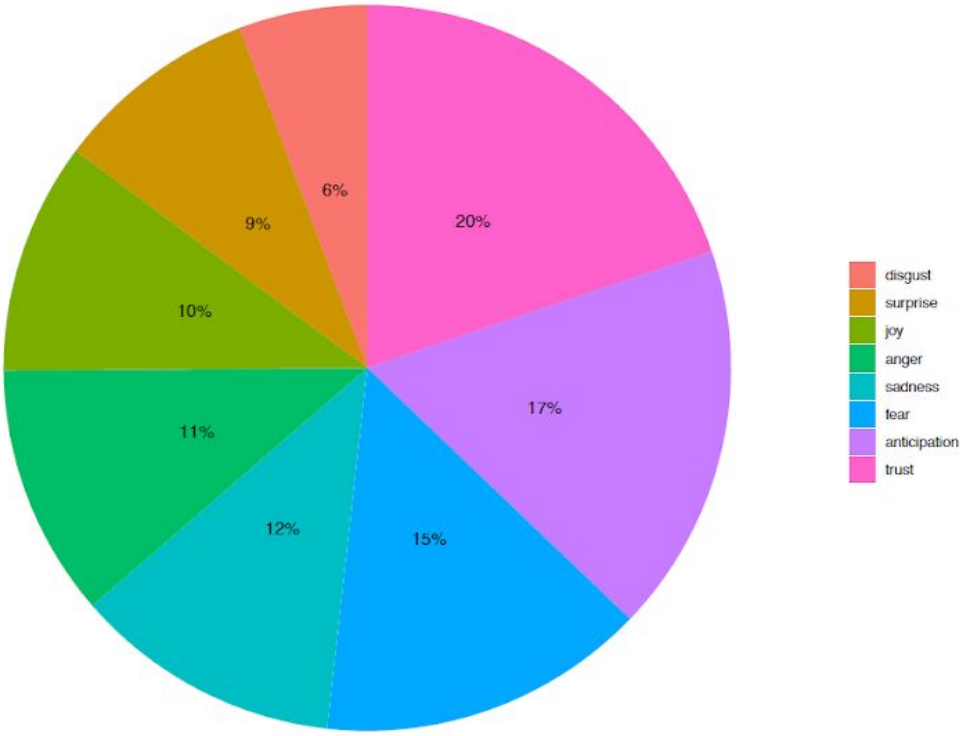
This section shows how sentiment is towards Brexit and Uncertainty. The Tweets contained in the dataset, analysed through Lexicoder Sentiment Analysis, give a negative score, which is maintained through the constituent members' subdivision of the dataset, with some slight differences (Northern Ireland being the "least negative"). Next up, we want to have another way of analysing sentiment, through NRC Sentiment Analysis.

3.4.2 NRC Sentiment Analysis

NRC analysis is more profound than the one we just did. From figure 40 we clearly see that the predominant sentiments are Trust (20%), Anticipation (17%), Fear (15%), Sadness (12%), Anger (11%), Joy (10%), Surprise (9%), Disgust (6%). We have to be

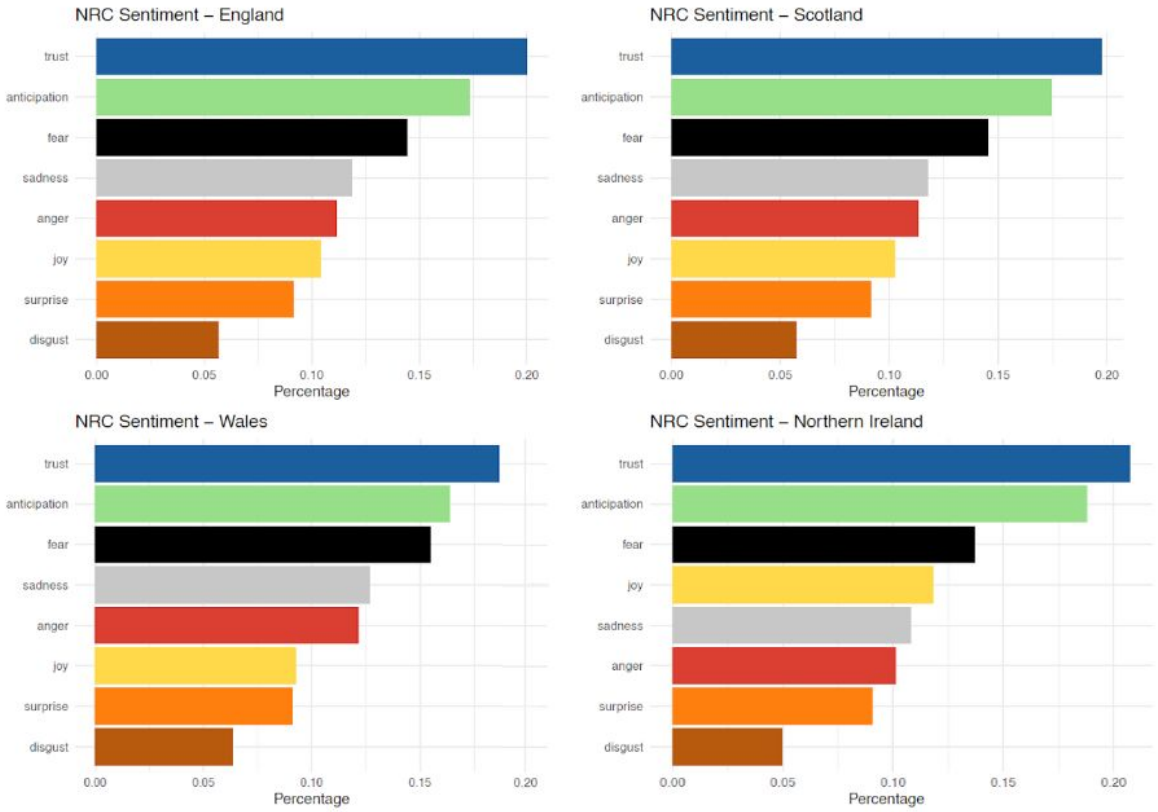
cautious, though, because we cannot explain whether the sentiment of trust has positive or negative valence. For example, the sentence "I trust Theresa May" refers to the“Trust” emotion, and it’s positive, but also the sentence “I don’t trust Theresa May” does,but with a negative connotation (absence of trust). What is interesting is that the main emotion is the one of Trust, whereas we would expect other emotions to be prevailing such as “fear” or “sadness”, since all these Tweets are related to uncertainty.

Figure 40: NRC Sentiment (UK)



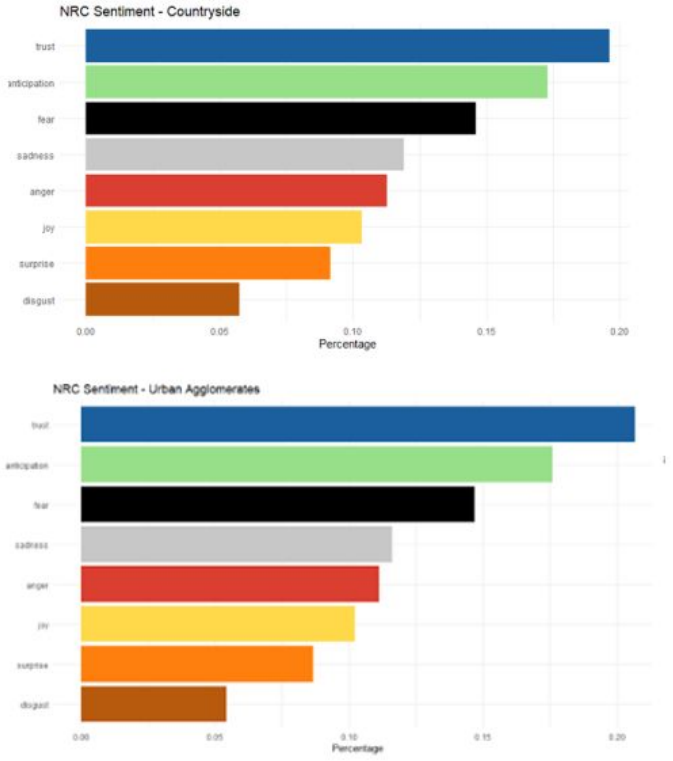
If we divide NRC sentiment by the four constituent countries of the UK we don’t find many differences, except for Northern Ireland, that seems to have a higher value of the “joy” sentiment (figure 41).

Figure 41: NRC Sentiment (divided by UK constituent countries)



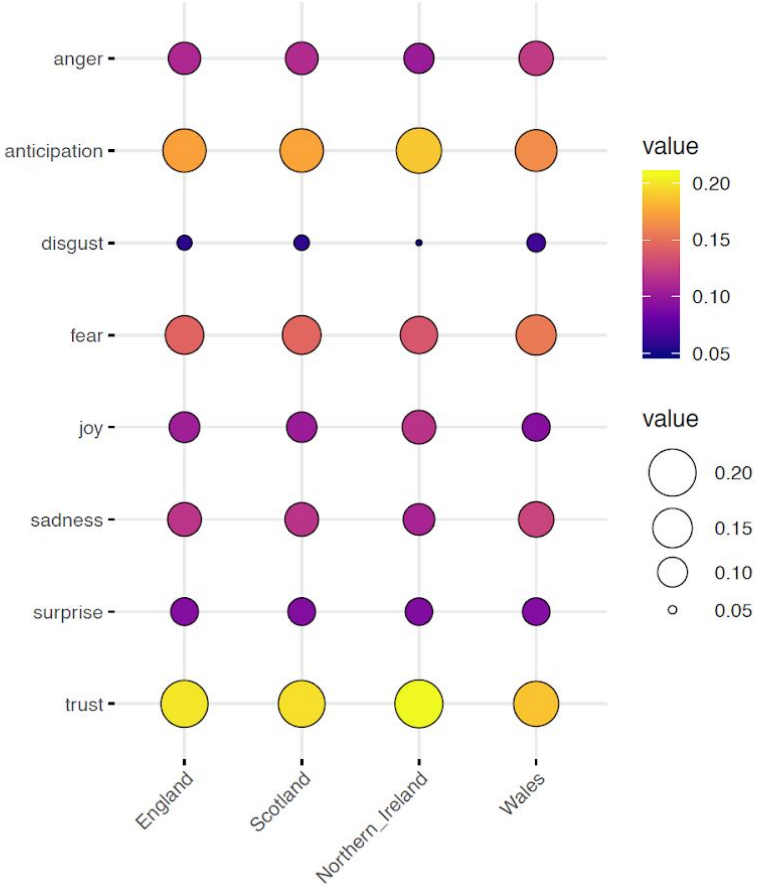
Further dividing the dataset into the other two subsets (urban agglomerates and countryside) does not show particular differences (figure 42). The only significant difference is in Northern Ireland, where joy sentiment shifts from penultimate place to fourth place. The other emotions, although with varying percentages across constituent countries, seem to stay on the same order. The same is true for Urban Agglomerates and Countryside division in figure 42, where apart from a slight variation in the "disgust" emotion (which appears to be slightly lower on Urban Agglomerates), they seem to be pretty much equal in terms of position.

Figure 42: NRC Sentiment (divided by Urban Agglomerates and Countryside)



The Balloon Plot in figure 43 simply but effectively highlights the main NRC sentiments that arise from the analysis. Color and dimension of the balloons are redundant in order to highlight differences in the results.

Figure 43: Balloon Plot of NRC Sentiment (divided by UK constituent countries)



This analysis helped us "colour" our dataset of tokens with 8 emotions, each assigned to one token, with the NRC criterion. We have seen that the emotion of trust is the most used out of all 8, followed by Anticipation, Fear, Sadness, Anger, Joy, Surprise, Disgust. When further subdividing the dataset into geographical areas, we see no particular difference, showing that there is consistency among different geographical areas. What is really curious is that the predominant emotion is trust, and not other emotions traditionally associated with uncertainty (such as fear or sadness).

4. Discussion

The results obtained through this analysis enlighten many interesting points regarding uncertainty in UK during the Brexit process and economic variables related to such uncertainty.

This Frequency Analysis helped us gain a general point of view of the dataset. We now understand that the most frequent words come from two macro categories, "political" and "economic" words. Some issues are peculiar to some Constituent Countries (Northern Ireland's Border issue, as well as the "Court" issue in Wales linked to "Continuity Bill"). A real surprise is the fact that @number10cat is the most quoted Twitter User that has been called into question by British people in Tweets regarding "Brexit and Uncertainty". This really is a proof of British humour. Finally, we are able to spot the most important "Remain" hashtag campaigns and their provenance (Scotland, Wales, Northern Ireland). Our aim of understanding how this condition of uncertainty deals with socio-economic variables is met particularly through taxonomies. In fact, the main point to be made is that through the use of taxonomies we are able to dissect the dataset in a more informative way for our socio-economic analysis with respect to the classical sentiment and lexical diversity analysis. The taxonomy on economic variables, for example, sheds light on what the public of Twitter considers to be the most important variables, namely "trade", "tax", "demand" and "domestic product". Moreover, Twitter's political scene seems to be dominated by the Conservative party, as they are the most quoted. The same is true when digging for the most quoted MPs, as Theresa May and other Conservatives are the most quoted politicians that are members of the parliament. We must keep in mind that this does not indicate whether the attention given to this party is positive or negative, it just means that it has a strong position in relation to uncertainty in a very popular social media like Twitter. Of course they are the most linked to uncertainty, since conservatives are the ones who mainly advocate for "Brexit". Another interesting result comes from the most quoted companies among the Top 50 by revenue: here we can see how "familiarity effect" might influence Twitter's perception of problems and of principal actors, but only up to a certain degree. In fact, SSE, the most quoted UK Company among the Biggest 50 by revenue, was at the

center of a great debate at the beginning of 2018, regarding a revolutionary energy project that should take advantage of strong linkages between EU and UK to take place, not the other way around. This has caused a lot of uncertainty, along with other examples mentioned above.

The Network analysis allowed us to understand how the different socio-economic variables are linked between them thanks to semantic networks that show nodes and edges of a network, and also some calculations to support the analysis, such as degree, betweenness, average clustering coefficient and modularity.

From a linguistic standpoint, we don't see remarkable differences in lexical diversity across geographical areas of interest for our research, with a couple of exceptions (Wales and Urban Agglomerates) that tend to deviate in a limited manner. Finally, the sentiment of people leans towards the negative side, as one should expect in times of uncertainty.

It is worth noting that for the majority of data that we have analysed, the results are not representatives of "causality" theories: many of these results serve just as a representation of how Twitter Users from the UK perceive and signal uncertainty, and how they link it to the variables we chose to perform the analysis on (our Taxonomies). In essence, the fact that people tend to quote "SSE" and "Reckitt and Benckiser" group is not an alert that they believe that uncertainty comes from these two companies; they are linking uncertainty during Brexit era to these two companies. In substance, we must not draw hasty conclusions and see causality where there is pure correlation.

More on that, we have seen how uncertainty can be described using some techniques that come from the world of data mining. In particular, the use of taxonomy is straightforward and flexible: whenever we want to analyse a dataset of fewer or more Tweets, over a shorter or longer timespan, we are allowed to do it within minutes.

What is even more interesting is that all these assumptions we are making are not the fruit of some journalistic reports or speculations. On the contrary, they are the result of direct observations of a very large sample of opinions (even though Twitter users are not perfectly representative in terms of distribution by age and schooling years) of the U.K. population, and it is very vast. In fact, a simple survey would hardly be able to let us collect such results as what we have done.

A final note has to be made regarding the four different analysis that we performed. While Sentiment and Lexical Diversity have some limitations and are a dry representation of the dataset, which is not so informative, it is thanks to the correct use of taxonomies that we are able to have a rich picture of how a share of the population link socio-economic variables to uncertainty, and in this case, an historical event such as Brexit.

5. Further Improvements

There are several further improvements still to be made to ensure progression in future research. First of all, it will be wiser to expand and better implement the taxonomies, for two main reasons. The first one is that knowledge is really hard to classify universally, and so the more people help do it, the better. The second reason is that this Master's Degree Thesis wants to be a starting point to prove the potential of this instrument, which can be exponentially improved when further implemented.

Furthermore, a geographical section of a broader area should be pursued. The dataset should contain Tweets that come from more diverse nations, such as African and European Countries. In this analysis the results of the geographical analysis have almost always confirmed the aggregate data. It could be a coincidence, but maybe trying a larger area, with different cultures and points of view could help capturing a substantial difference.

Moreover, the taxonomy analysis could be transformed also into a time series analysis. Researchers could verify whether people link the same socio-economic variables to uncertainty during time, and investigate on why and how changes in perception happen.

Appendix

Full Taxonomies

Since the full Taxonomies would take up more than 50 pages of this thesis, we leave a link to access them online on Google Drive: http://bit.ly/brexituncertainty.
The Taxonomies are available both in .CAT and .txt file format, so that they are easily accessible to all readers, technical and non-technical.

Complete List of R Packages used for the Analysis

Here there is the full list of packages used for each analysis, in order to give credit to the authors and show how the analysis was done. Complete citation will be in the end of the Bibliography.

Data Extraction and Data Pre-Processing

quanteda, quanteda.corpora, tidyverse

Taxonomies

quanteda.dictionaries, networkD3, data.tree, rjson

Frequency Analysis

quanteda

Taxonomy Frequency Analysis

quanteda

Taxonomy Network Analysis

quanteda, igraph, qgraph, data.table

Lexical Diversity Analysis

quanteda

Sentiment Analysis

quanteda, raster, sp, dplyr, plyr, ggplot2, httr, ROAuth, stringr, twitteR, syuzhet

Bibliography

M. Warglien, (2019) “Research fellowship on ‘A web observatory of uncertainty’ - Università Ca' Foscari Venezia” Prot. n. 17687 rep. 291

C. Santagiustina, (2018) “Talking About Uncertainty”

S. R. Baker, N. Bloom, and S. J. Davisc, (2016) “Measuring Economic Policy Uncertainty”

R. Rogers (2013) “Debanalizing Twitter: The Transformation of an Object of Study”

E. Borra and B. Rieder (2014), "Programmed method: developing a toolset for capturing and analyzing tweets," Aslib Journal of Information Management, Vol. 66 Iss: 3, pp.262 - 278. http://dx.doi.org/10.1108/AJIM-09-2013-0094

N. F. Noy, D. L. McGuinnes (2001), “Ontology Development 101: A Guide to Creating Your First Ontology”

D. Nettleton (2014), Commercial Data Mining

P. W. Foltz (2001), Semantic Processing: Statistical Approaches

J. Borge-Holthoefer, A. Arenas (2010), Semantic Networks: Structure and Dynamics

T. Kwartler (2017), “Text Mining in Practice with R”

HM Government (2017), “The United Kingdom's exit from and new partnership with the European Union”

L. Laineste (2014). "National and Ethnic Differences". In Attardo, Salvatore (ed.). Encyclopedia of Humor Studies. SAGE Publications. pp. 541–542.

G. Tetlow and A. Stojanovic (2018), “Understanding the economic impact of Brexit”

W. Wright, C. Benson & E.F. Hamre (2019). “The New Financial Brexitometer”, New Financial

G. Huberman (2015). “Familiarity Breeds Investment”, The Review of Financial Studies

R. Koizumi (2012), Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, A. Matsuo (2018). “quanteda: An R package for the quantitative analysis of textual data.

K. Benoit and S. Müller (2019). quanteda.dictionaries: Dictionaries for Text Analysis and

Associated Utilities. R package version 0.22.

K. Benoit (2019). quanteda.corpora: A collection of corpora for quanteda. R package version 0.87. http://github.com/quanteda/quanteda.corpora

H. Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

J.J. Allaire, C. Gandrud, K. Russell and CJ Yetman (2017). networkD3: D3 JavaScript Network Graphs from R. R package version 0.4. https://CRAN.R-project.org/package=networkD3

C. Glur (2018). data.tree: General Purpose Hierarchical Data Structure. R package version 0.7.8. https://CRAN.R-project.org/package=data.tree

A. Couture-Beil (2018). rjson: JSON for R. R package version 0.2.20. https://CRAN.R-project.org/package=rjson

G. Csardi, T. Nepusz: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

Sacha Epskamp, Angelique O. J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, Denny Borsboom (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. Journal of Statistical Software, 48(4), 1-18. URL http://www.jstatsoft.org/v48/i04/.

M. Dowle and A. Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.2. https://CRAN.R-project.org/package=data.table

R. J. Hijmans (2019). raster: Geographic Data Analysis and Modeling. R package version 3.0-2. https://CRAN.R-project.org/package=raster

Pebesma, E.J., R.S. Bivand, (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

H. Wickham, R. François, L. Henry and K. Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

H. Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

H. Wickham (2019). httr: Tools for Working with URLs and HTTP. R package version 1.4.1. https://CRAN.R-project.org/package=httr

J. Gentry and D. Temple Lang (2015). ROAuth: R Interface For OAuth. R package version 0.9.6. https://CRAN.R-project.org/package=ROAuth

H. Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

J. Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR

M.L. Jockers (2015). _Syuzhet: Extract Sentiment and Plot Arcs from Text_. <URL: https://github.com/mjockers/syuzhet>.

Sitography

<https://dictionary.cambridge.org/dictionary/english/uncertainty>

<https://www.pwc.com/gx/en/issues/economy/global-economy-watch/how-does-uncertainty-impact-economic-activity.html>

<https://www.policyuncertainty.com/index.html>

<https://www.economist.com/finance-and-economics/2019/08/15/the-chilling-economic-effects-of-brexit-uncertainty-are-intensifying>

<https://www.theguardian.com/business/2019/aug/09/uk-economy-contracts-on-back-of-brexit-uncertainty>

<https://developer.twitter.com/en/docs.html>

<https://www.techopedia.com/definition/13698/tokenization>

<https://www.bbc.com/news/uk-wales-politics-43795158>

<https://www.bbc.com/news/uk-politics-48826360>

<http://mentalfloss.com/article/592523/larry-the-cat-uk-chief-mouser-facts>

<https://www.theguardian.com/media/2018/jan/17/fbpe-what-is-pro-eu-hashtag-spreading-across-social-media>

<https://www.bbc.com/news/uk-47213842>

<https://www.theguardian.com/business/2019/aug/09/uk-economy-contracts-on-back-of-brexit-uncertainty>

<https://sse.com/newsandviews/allarticles/2018/02/when-it-comes-to-energy-collaboration-with-the-eu-is-imperative/>

<https://www.taylorwessing.com/download/article-media-entertainment-after-brexit.html>