STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic year 2020–2021

Can We Predict Business Cycles With Natural Language Processing?

Albert Flak (41642)

Abstract: For centuries, many economists have attempted to solve the puzzle of business cycles; what explains them, and is it possible to predict them? The great amount of electronically available textual data, together with the recent advancements in unsupervised natural language processing are bound to offer new ways of analysing these perennial questions. Inspired with the emergence of narrative economics as a new field of economic analysis, I propose a novel research design to computationally extract business cycle sentiment from textual data on economic expectations. Firstly, using word vectorisation and embedding techniques on a large corpus of news articles on economic boom and bust, the computer learns to identify expansionary and contractionary sentiment. Secondly, Latent Dirichlet Allocation is used to categorise economic expectation news into topics, and Shannon's Entropy of the topic distribution enumerates the broadness of discourse. The insights from both approaches are used to construct two indices – Relative Sentiment and Narrative Consensus Index – to foresee business cycle turning points and predict realisations of U.S. Gross Domestic Product growth. Correlations between current news content and future realisations of U.S. GDP growth as well as several other key macroeconomic time series are established. The predictive power of the indices is evaluated. The Relative Sentiment Index is found to be strongly related to current and short-term future macroeconomic outcomes, and to identify NBER U.S. business cycle turning points with a lead of up to five months. The research method of the thesis offers inspiration for further computational narrative research in macroeconomics, and the conclusions provide preliminary evidence of the potential relevance of economic storytelling for future macroeconomic outcomes.

JEL: C53, D83, E32, E37, E71, G01
Keywords: Business Cycles, Economic Narratives, Natural Language Processing (NLP), Latent Dirichlet Allocation (LDA), Word Embeddings, Early-Warning Systems (EWS)

Supervisor: Andreea Enache
Examiner: Kelly Ragan
Discussant: Ingrid Löfman
Date submitted: December 5, 2020
Date examined: January 8, 2021

Contents

List of Figures

List of Tables

1 Introduction

Narrative economics, the study of the viral spread of popular narratives that affect economic behaviour, can improve our ability to anticipate and prepare for economic events.

— Shiller (2019, p. 3), *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*

The thesis title asks a comprehensive, open-ended question: Can We Predict Business Cycles With Natural Language Processing? Robert Shiller, the Nobel-Prize winning economist whose work on narrative economics proclaims that "stories matter" for economic outcomes, would most likely want us to answer the question with a definitive and resounding Yes. In fact, the notion that what beliefs people hold about the workings of the economy, particularly about its future evolution, and how these beliefs are synthesised in popular talk matters for macroeconomic outcomes is by no means restricted to Shiller. The notion can be found repeatedly, manifested in different ways and contexts in the works of great economic thinkers such as John Maynard Keynes and Hyman Minsky, or of contemporary scholars such as Andrei Schleifer. For example, in his General Theory, Keynes (1936) describes a beauty contest where the outcome of such a market coordination game is given by every market participant forming expectations subject to their expectations of others' expectations. This clearly necessitates that every market participant first had to formulate a story, perhaps a convincing narrative, about others' expectations so that he can build his own expectations. Can we capture such a guessing game with textual analysis? A couple of years later, Minsky (1992) formulated his financial instability hypothesis – another example of how economic narratives could matter. Minsky postulated that an economic expansion would inevitably lead to excessive lending and, ultimately, to the instability of the financial system whose key players will at some point have to collectively realise that their beliefs about market fundamentals have been erroneous. At that point, the conviction changes, likely towards greater pessimism, and an economic contraction – perpetuated by a contraction in lending – *follows*. Can we capture Minsky's moment – the business cycle turning point – with textual analysis? Nowadays, systematically incorrect beliefs seem to be one of the promising explanations of crisis emergence. Gennaioli and Shleifer (2020) describe a model of rational expectation formation that draws on overconfidence and extrapolation to explain how systematically incorrect beliefs emerge, and later collapse when the convictions become unsustainable under the face of the fundamental reality. They present empirical evidence from surveys of market expectations that beliefs can rationally align in an objectively incorrect equilibrium and are persistently correlated over different sources. Can we spot alignment in expectations with textual analysis?

Nonetheless, the arguments of these many economists, and by implication, the claim that popular talk – the stories we tell – matter for economic outcomes, is anything but a tenet of standard economic teaching. Despite this apparent discord, most people would probably not dispute that the above economists' arguments sound intuitively plausible. Whereas their work features many potential theoretical explanations for *why* and *how* economic storytelling can matter, the plausible view that stories matter seems to be empirically under-analysed, and very little empirical evidence of this connection exists to date. This is especially true for the relationship between economic stories and macroeconomic variables, such as the Gross Domestic Product. The reason for this lack of empirical evidence is most likely the fact that it is not at all straightforward how a researcher should go about analysing economic storytelling in textual data.

It is natural to expect that many of these economic beliefs about the future, if they are ever written anywhere explicitly, will be stated in textual accounts such as newspapers or (social) media where opinions about economic concepts and events are discussed in high frequency. The thesis collects and analyses 30 years of economic storytelling in newspaper articles with particular relevance for business cycles and macroeconomic expectations. The importance of text for economics makes it worthwhile to develop methods to analyse textual data sources computationally – a point emphasised i.a. by Gentzkow, Kelly, and Taddy (2019). The field of natural language processing (NLP) deals with precisely that – it offers tools to analyse human language computationally. Robert Shiller, in his work on economic storytelling in Shiller (2017) or Shiller (2019) often emphasises that narrative economics should draw on the knowledge from different academic spheres – notably the epidemiology. In this thesis, I shall demonstrate that it is perhaps also the wisdom of linguists and computational linguists in particular that is needed to develop the field of narrative economics further. More broadly, the purpose of this thesis is

to apply a novel class of natural language processing techniques to analyse their usefulness for mathematically capturing economic beliefs and narratives from textual data. Since some of the techniques have never been applied in the field of economics before, the work here is highly explorative – its results are a starting point for further research, subject to several conceptual and computational limitations, and could at times strike the reader as inducing many new questions. Nevertheless, the aim is to meaningfully contribute towards the literature on business cycle emergence, narrative economics and economic literature using natural language processing methods, both in terms of methodology and conclusions.

To identify and shed new light on the connection between text and future macroeconomic outcomes, two purely text-based indices are created. Firstly, the thesis develops a methodological framework for inferring and mathematically representing business cycle narratives as collections of words found probabilistically related to business cycle keywords in news reporting. This semantic relationship between business cycle keywords and other words is inferred in applying the GloVe technique of Pennington, Socher, and Manning (2014) to represent word meaning in a compact vector space that compresses linguistic patterns in business cycle news reporting and allows meaningful algebraic operations on the inferred word vectors. The relative incidence of the words closely associated with business cycle keywords over time is thereafter examined in a corpus of news articles on economic expectations. This relative incidence of the business cycle words found representative of different business cycle stages pins down two sentiments – a contractionary and an expansionary sentiment – whose difference is examined to predict realisations of U.S. GDP growth. Secondly, a measure of narrative consensus is created to proxy the level of agreement based on text from news articles on economic expectations. This is achieved by structuring articles' content into several categories defined by their probability distribution over representative words via the Latent Dirichlet Allocation technique introduced in Blei, Ng, and Jordan (2003). Later, Shannon's entropy of this categorical structure is calculated to periodically determine the broadness of discourse in the news.

The main task of the thesis will be the construction of these purely text-based indices to demonstrate that they could exhibit the following three main properties. Firstly, they display a stable and econometrically evidenced relationship to a measure of GDP. For comparison, there already exist several purely text-based indices which have been found to exhibit relationships with current GDP. Examples of this are the Economic Policy Uncertainty Index of Baker, Bloom, and Davis (2016) or the GDP-coincident indices constructed by Larsen and Thorsrud (2019b), Larsen and Thorsrud (2019a) and Thorsrud (2020). Secondly, they are able to (at least in the very short-term) predict the evolution of GDP. To the best of my knowledge, there is no NLP-based economic measure that would attempt exactly this. There is, however, the recent work by Shapiro, Sudhof, and Wilson (2017) who have attempted to predict measures of consumer confidence with NLP techniques. Thirdly, and finally, the indices could be able to *predictively* capture the cyclical nature of the macroeconomy – the business cycle turning points – and as such serve as a basic early crisis warning system. Using an analogy, the ideal index should be a text-based counterpart of the inverted yield curve. Inversion of the domestic yield curve has been widely documented to predict emerging crises (see, e.g., the standard approach of Rudebusch and Williams (2009) or the recent machine-learning approach of Bluwstein et al. (2020) showing this relationship). If the research could credibly demonstrate the existence of a text-based index with early crisis prediction capabilities, the relationship between economic storytelling and *future* macroeconomic outcomes could be empirically established. It should be stressed that no causal inference of this relationship was yet found. The thesis merely aims to establish the existence of correlations and predictive power of written text towards the economy. It hopes to motivate further research using similar methods and evaluating related research questions.

I shall attempt the creation of these text-based, narrative-capturing macroeconomic indices in the following steps. In section 2, I review the bulk of different strands of economic literature that the research question derives from, and the thesis is connected to. In particular, I attempt to connect the research question to works in the field of business cycle analysis, narrative economics, and to other works in macroeconomics which have used natural language processing. In section 3, I introduce and discuss the data collection approach, enriched with descriptive statistics of the final corpora. In section 4, I offer a thorough discussion of the classes of natural language processing techniques used in the course of this thesis and write about several crucial aspects of the nature by which computers learn about human language and how they make inferences based upon their learning. The section progresses from discussing data pre-processing techniques to common natural language processing algorithms and goes on to develop a coherent and justifiable analytical framework for the narrative indices. The

construction of the indices is elaborated upon. The time series and econometric tools used to evaluate the success of the two indices in answering the research question, and in their success to achieve the goals outlined above, are discussed and explained. Section 5 graphically presents the two indices, describes and discusses their interesting properties. Evaluation results are presented and discussed. Section 6 offers a discussion on limitations, further potential extensions and adjustments to my study. Finally, Section 7 concludes, and the Appendices A–Q feature many accompanying graphs, econometric results and methodological intuition.

2 Literature Review

The combination of business cycles, narratives, and natural language processing touches upon many, seemingly distant fields of research. Combined with the relative novelty of the research questions and the method, a comprehensive literature review would be rather lengthy. Nevertheless, the attempt is to provide a broad context, and highlight the reasoning, to further strengthen the case for the thesis' research questions. Firstly, the classic business cycle literature from which the thesis draws the most inspiration will be examined. Secondly, the role of narratives in the theory of economic fluctuations is outlined. Thirdly, although only recently a part of economic research, the use of textual analysis and machine learning techniques in the economic literature is reviewed. Fourthly, the research most closely related to the thesis, such as combining insights from all the previous three categories, is discussed.

2.1 Economic Theories of Boom and Bust

It is curious to notice the variety of the explanations offered by commercial writers concerning the cause of the present state of trade... But why do we beat about the bush when all that is needed is half-a-dozen of Pouillet's pyrheliometers ... to determine directly the heating power of the sun? ... From that sun, which is truly "of this great world both eye and soul", we derive our strength and our weakness, our success and our failure, our elation in commercial mania, and our despondency and ruin in commercial collapse.

— Jevons (1878), *Commercial Crises and Sun-Spots*

Why do major economic crises and bubbles keep happening? To answer this perennial question, one needs a theory that goes beyond that of market equilibrium. As far as one accepts the idea that the economy can, for a substantial period of time, find itself away from a hypothetical equilibrium, a theory of boom and bust is possible. Throughout much of history, economists either gathered around the idea that business cycles are a recurring, expectable characteristic pattern of the economy, or that they result from series of stochastic shocks. The latter is commonly noted to derive from Slutzky (1937), who has demonstrated that when summing randomly generated time series, in his case, the results of a Russian lottery, cyclical patterns similar to those of business cycles easily emerge[1]. Modern macroeconomic theory, based on the new neoclassical synthesis[2] and real business cycle theory grounded in Lucas (1972), Lucas (1977), and Kydland and Prescott (1982), continues to rely heavily on this logic. As such, modern macroeconomic modelling mostly features exogenously driven stochastic shocks to different variables that either as a sum or individually create cyclical patterns in economic growth via a micro-founded propagation mechanism. After the shock, the variables tend towards their natural equilibrium. An expansion on itself thus does *not* effect a contraction; both expansion and contraction are results of different stochastic shocks. This modern assumption does *not* imply that one cannot find economic explanations for cycles, but it *does* imply that it is fruitless to predict them. If business cycles result from a *random summation* of *random* events, not only would a causal understanding of crises and booms not help us to predict them, *any* systematic prediction (over a sufficiently long time period) whether causal or not would be fruitless. The question in the title of the thesis

[1] Another economist demonstrating similar mechanisms, albeit in a different fashion was Udny Yule. In Yule (1927), he highlights the "nonsense correlations" in economics and shows how real disturbances in the form of a white noise hitting a pendulum could produce cyclical patterns resembling those of economic variables.

[2] The reader is kindly encouraged to refer to the expositions in Goodfriend and King (1997) and Woodford (2009) for a summary review of the theoretical pillars of the modern mainstream macroeconomics.

could not be answered positively.

To give the endeavour any chance at succeeding, we need to look beyond the mainstream macroeconomics. Particularly towards the economists that have claimed the cycles to be a recurrent, inherent characteristic of the economy. For the earliest clues, one can look as far back as the work of Juglar (1862). Clément Juglar developed a much-appraised theory of business cycles, substantiated with ample contemporary empirical evidence, that led Schumpeter (1954) to consider him the founder of modern business cycle theory (Legrand & Hagemann, 2007). The central proposition of his theory can be summarised as an inherent instability of the market economy that, through its inherent proneness towards future financial and business speculation, tends towards an economically unsustainable state of prosperity that has to find its end. The cycle, as Juglar sees it, is driven by the interplay between various market prices, expansion and contraction of credit, and the aggregation of beliefs about their future evolution. It is perhaps this link between financial mechanisms and real economic activity via agent's collective beliefs that is so convincing, from today's viewpoint, and what prompted Joseph Schumpeter to so highly regard the early economist as to incorporate many of Juglar's basic principles into his theory of business cycles in Schumpeter and Opie (1934) and Schumpeter (1939). If such systematic reversions in collective beliefs can be cleverly captured with natural language processing, the resulting insight could be able to predict business cycle turning points.

The emphasis on different theoretical foundations of expectation formation is perhaps the main aspect along which much of the heterodox theories of the sources of business cycles differ from the prevalent mainstream. Instead of the emphasis on rational expectation hypothesis based on Lucas (1972), alternative approaches to the mechanism of future expectations formation[3] are seen crucial to explain radical changes in economic environments and the crisis tipping points. A notable example, and where the thesis draws the most inspiration, is in the financial and behavioural theories of business cycles as embodied in the thoughts of Frank H. Knight, Hyman Minsky and Robert Shiller. Arguably, many of the past financial and ensuing economic crises have begun with a situation that could be broadly characterised as a "Minsky moment"[4]. Common to these situations is a sudden correction in market beliefs related to the perception of risk about important market fundamentals. As the perception of risk changes, so do market expectations and valuations. This sudden realisation and the following chain of corrections can cause a vicious circle of events, new information, and further belief updating that has the potential to result in serious macroeconomic implications such as a growth slowdown or a recession.

How is it possible that beliefs change so abruptly? Intuitively, the relevant beliefs about the future are bound to be incorrect in some sense, and new information could cause market actors to update their beliefs with that new information. However, even this basic intuition, that a crisis has to be caused by market agents updating their beliefs with new information has been disputed on the grounds of historical evidence. In his book, Shiller (2015) reviews the context of the crises of the past centuries. He finds that neither the Crash of 1929 nor the Crash of 1987, can be traced back to specific market information. The author confirms the latter by his survey of market participants in Shiller (1987). His findings are clearly in conflict with any notion of business cycles in the sense of the new neoclassical synthesis. If these business cycle turning points cannot be traced to a specific shock, much less can any frictions in the real economy result in them, and the theory does not provide any immediately credible explanation of the sources of the ensuing macroeconomic consequences. If agents understand the workings of the economy, i.e., understand the grand "model of the economy", how come that such a major correction can happen without a distinguishable cause?

The most recent work of Robert Shiller (e.g., Shiller, 2017, 2019) seems to strongly suggest that a crisis can occur simply on the basis of a self-fulfilling prophecy. Perhaps then, a crisis does not even prerequisite a relevant, new and unexpected piece of information to come to light, but merely some mechanisms behind human psychology could be at play. Bénabou and Tirole (2016) explore the process of belief formation and suggest various potential mechanisms by which a belief of a market participant about an economic fundamental could be (predictably) wrong. According to Bénabou and Tirole (2016), an individual faces a constant trade-off between accuracy and

[3] Overview of several alternative approaches is offered by Woodford (2013) and a brief discussion of the evolution of macroeconomic paradigm after the financial crisis in Landmann (2014).

[4] Comprehensive review of Minsky's Financial Instability Theory is beyond the scope of this review. The reader is kindly referred to consult Minsky (1992). According to Lahart (2007), the term Minsky moment was first used by Paul McCulley, then managing director of PIMCO.

desirability while forming beliefs. As a result, in certain situations, it is rational for an individual to employ self-deception strategies, such as strategic ignorance, reality denial and self-signalling, in order to form the most useful, yet wrong, beliefs. In the context of the emergence of crises, this could bear two critical implications. Firstly, market agents could be rationally overconfident, overestimating their abilities to hold correct beliefs. For example, it has been shown that agents make predictable forecast errors, mostly as a result of simple extrapolation, and that they underestimate mean-reversion in financial time-series. Examples of the evidence include Bordalo, Gennaioli, and Shleifer (2018), Greenwood and Shleifer (2014) or Gennaioli, Ma, and Shleifer (2015). Secondly, market agents could be rationally over-optimistic about future prospects, such as by disregarding negative information. This proposition could be the essence of the neglect of tail risk hypothesis of many crises (e.g., Taleb, 2007). If we then could identify the moment when market participants' state of mind changes from overconfidence and overoptimism towards greater skepsis, perhaps we could capture the essence of Minsky's moment.

In light of all this research, the essential task of textual analysis of business cycles seems to lie in developing an understanding of how market actors think about booms and busts. For example, if agents associate a certain vocabulary strongly with the emergence of a crisis, the mere shift towards this vocabulary could be an indication of problems ahead, could make economic agents more attentive, and ultimately more likely to update their beliefs *strongly*. To better understand what states of the world, economy and financial marketplace, the agents see as indicative of a crisis or an exceptional enthusiasm, the thesis aims to analyse news reporting over longer periods of time. The key is to identify concepts – vocabulary – that relatively often occurs when news reporters write about economic crises, recessions, depressions, as compared to when booms, prosperity, and expansions are discussed. The ultimate challenge rests in the identification of such vocabulary that is all relevant, potentially indicative, and characteristic of the meaning of particular business cycle stage, providing a proxy for the prevalent beliefs and sentiments among market actors. An analytical, evidence-based overview of such associative vocabulary among agents could potentially help economists and market actors to understand where the risks for a market correction a la Minsky's moment amass before it needs to happen. Capturing this process is the primary underlying motivation of the thesis. How to know what to look for? The review provided in the next part should offer several clues.

2.2 Narrative Theory and Economic Narratives

This is what fools people: a man is always a teller of tales, he lives surrounded by his stories and the stories of others, he sees everything that happens to him through them; and he tries to live his own life as if he were telling a story.

— Sartre (1964, p. 39), *Nausea*

Is textual data a valuable source of empirical evidence in economics? It better be, if we hope to find a positive answer to the question posed in the thesis' title. In Shiller (2017), the financial economist offers a vision of a new field of economic study, perhaps a theory on its own. The insight is simple and clear – stories matter. *What* we tell to each other, to ourselves, and *how* we do it, is not only able to explain the past, but could also give us clues as to what to expect in the future. There are many key questions to be answered: What constitutes a narrative? When is a narrative economically, and especially, *macro*economically relevant? How and why do economic narratives evolve? What is their relationship to beliefs, to expectations about the future, and to economic fluctuations? Is it possible to capture and quantify them? Regrettably, economics does not seem to be in abundance of answers to these, and many related[5], questions.

On first sight, the word story and the word narrative are very similar, if not the same. However, the two are bound to be profoundly different, especially if one considers the way the word narrative is used by Robert Shiller or many sociologists that work with narrative analysis. Drawing on the review of the concept by Franzosi (1998), for a body of text to contain a narrative, it needs to recount a story (sequence of events) in a linear (sequential),

[5]The only extensive study of what could be considered macroeconomic narratives beyond the definition of Shiller (2017) is perhaps the narrative approach to the analysis of monetary policy effects proposed by Romer and Romer (1989) and continued in Romer and Romer (2004). Their approach has later been employed in many different areas to identify shocks such as in government spending (Ramey & Shapiro, 1998; Romer & Romer, 2010). These narrative-driven exogenous shocks have also been integrated into the VAR and SVAR models of monetary policy effects in Stock and Watson (2012) and Antolín-Díaz and Rubio-Ramírez (2018).

organised, and connected fashion, so as to be logically comprehensible to the narrator. The connections between words and sentences are particularly important in this regard, which will be the essence of the focus on the probabilistic description of word context in the analysis later on. As the author writes, a story needs both, a turn of fortunes, which could, in our case, be a new piece of information in the news, and a temporal ordering. The story is then transformed into a narrative when we recount or remember it. The Oxford English Dictionary gives two relevant definitions of the term narrative. First mirrors the preceding discussion exactly, defining narrative as: "An account of a series of events, facts, etc., given in order and with the establishing of connections between them". This definition begs the question – what kind of account of events? Any account? An objective account? The first definition is very vague. The second definition offered by the dictionary explains a narrative as "a story or representation used to give an explanatory or justificatory account of a society, period, etc."[6] (Oxford English Dictionary, 2003). This definition certainly makes the concept more precise. By giving to a story a justification or an explanation, one needs to draw on own knowledge about the world and possibly even on one's beliefs, values and ideologies. Therefore, narratives, whether economic or not, often have an emotional and subjective bearing that is of non-trivial importance for one's perception of the world, and as a consequence, are involved in the process of forming expectations about the future.

Therefore, narratives could influence one's decision making, and so influence economic outcomes. This ultimate implication is, in essence, what Robert Shiller propagates in Shiller (2017). He exemplifies this logic, inter alia, by the proven existence of both the framing effect and representativeness heuristics of Daniel Kahneman and Amos Tversky which he sees as symptoms of the economic narratives of humans – the *homo narrans*[7]. Convincing and contagious stories can serve as a frame and a reference point that shapes the logical reasoning of humans. Under what conditions should stories then be considered relevant for human decision-making and macroeconomic outcomes? There does not appear to exist practically any economic theory to answer this question. Shiller (2019), however, offers several propositions of narrative economics which he holds for particularly probable. Several of them have important implications for the subsequent analysis, particularly the following:

- *"Important economic narratives may comprise a very small percentage of popular talk"* (Shiller, 2019, p. 89)
  If true, any identification strategy cannot rely on simple counting methods – we cannot search for the 'most frequent storytelling'. Proportionalities and relative probabilities will thus be crucial in the analysis below.

- *"Narrative constellations have more impact than any one narrative"* (Shiller, 2019, p. 92)
  If true, the narrative analysis should rely on global, and not only on local statistics. If we try to identify patterns, these patterns should aim to be general. The machine learning employed later on relies on *global* statistics of the textual database.

- *"The economic impact of narratives may change through time"* (Shiller, 2019, p.93)
  If true, a narrative can lay dormant for a time. Optimally then, it could be essential to understand when a narrative will be acted upon, and when not. This proposition is partly a motivation for the focus on alignment of economic expectations in the analysis below.

- *"Narratives thrive on attachment: human interest, identity, and patriotism"* (Shiller, 2019, p.100)
  If true, narratives have a non-trivial sentimental, or emotional, value for the agent. The indices constructed below do explicitly accommodate this view.

[6] This definition of the word narrative is noted to derive from the structuralist system of thought in literature and anthropology. In particular, it posits that the ultimate ontological perception of humans is the direct result of the material structure of our world. Narratives in this sense embody the human perception of this materialist system. The more recent post-structuralist tradition rejects such notion of 'objective' understanding. It emphasises, in turn, the uncertainty involved in learning and the inability of humans to objectively grasp this material system of our world. It can prove difficult to grasp the true meaning of the word narrative as used by Shiller, or when the concept is used in reference to narrative analysis and narratology, without studying these philosophical sources of the concept. The reader is kindly referred to D. Palmer (2007) for a review. In the age of fake news and post-truth, it does not seem too difficult to believe that narratives in this sense could drive major events.

[7] For an explanation of the framing effect, see the work of Tversky and Kahneman (1981) and of the representativeness heuristic, kindly see Kahneman and Tversky (1972). *Homo narrans* or *homo narrator* or *homo narrativus* are all termini technici that have been used by Fisher (1984), Gould (1994) and Ferrand and Weil (2001) respectively to describe the human as a naturally and evolutionary narrative-based being.

The definition of the word narrative and Shiller's propositions for identifying relevant macroeconomic narratives offer clues that could help the goal of this thesis. Still, a key question remains: How can we proxy, or at least approximately, capture and quantify narratives? The most basic proposition, used to a great extent by Shiller, is tracking N-gram incidence. N-grams are collections of consecutive terms such as the bigram 'central bank' or the trigram 'asset-backed security'[8]. These will play a pivotal role in the analysis later on when business cycle sentiments are described with collections of unigrams and bigrams. Albeit the use of N-grams provides an example of a proxy of the evolution of narratives – as long as one could agree to call these N-grams narratives – it is only a single and lone attempt in the economic literature. In the field of sociology, Franzosi (2010) has attempted to provide a methodology to computationally extract narratives from a corpus of text. Building on the structuralist tradition, the author identifies three building blocks of each narrative – subject, action and object – that are commonly referred to as semantic triples. Building on his work, Sudhahar, De Fazio, Franzosi, and Cristianini (2015) have written on the automatic computational extraction, and especially on the network and graphic representations, of these triples. These authors draw on an area of natural language processing called part-of-speech tagging and syntactic parsing to identify the proposed semantic triples, which would be an exciting endeavour but will not be attempted here.

In the light of all this work, the grand question of how can we capture and quantify narratives relevant for business cycles, that would have significantly helped the achievement of this thesis' goals, remains unanswered in the economic literature. Without unambiguous and stable conditions as to the definition of a macroeconomically important narrative, it might be impossible to perform any conclusive computational analysis of economic narratives in written text. Therefore, this thesis does never claim that any authentic economic narratives in the sense of Robert Shiller could be uncovered in what follows. The identifying assumption for a macroeconomically relevant narrative in this thesis will be a substantial relative frequency in the co-occurrence of these candidate narrative words around specific business cycle keywords.

2.3 Text Mining and Natural Language Processing in Economics

But there is another message I want to tell you. Within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough.

— Draghi (2012), *Speech at the Global Investment Conference in London, 26 July 2012*

How common is it then to find textual analysis[9] in (macro)economic research? Not much. The early adopters of these computational techniques in macroeconomics can be found chiefly among central bank staff and monetary policy analysts. A summary example and beginner's handbook is Bholat, Hans, Santos, and Schonhardt-Bailey (2015). Another exposition can be found in Gentzkow et al. (2019). Beyond monetary policy, the other most significant strand of research utilising NLP deals with economic *policy* uncertainty. The originators of the method are Baker et al. (2016). They are the authors of the Economic Policy Uncertainty Index (EPU) which periodically counts news articles that fulfil certain criteria: articles that include at least one word from each of the three categories – pertaining to the economy, to uncertainty and economic policy. The indices are nicely shown to significantly correlate with measures of (option-implied) stock price volatility, measures of company investment, employment growth and industrial production. Their conclusions seem plausible insofar as higher levels of policy uncertainty negatively impact future economic outlook and vice versa.

[8] Figures 19 and 20 in Appendix A plots N-gram incidence of probable causes of the latest (2007–2008) financial crisis in a large corpus of English-speaking books. Interestingly, all of the examined concepts reach their local maxima in the period of the recent financial crisis and follow a roughly hump-shaped pattern. Moreover, many of these concepts have been increasing unprecedentedly on importance even *before* the crisis ensued. However, from the viewpoint of a pre-crisis economist, it is impossible to know what concepts to watch out for. Therefore the usefulness of such analysis for predictive purposes is rather weak.

[9] Textual analysis refers to both text mining and natural language processing. These two concepts have surprisingly vague and non-exclusive definitions. As a rule of thumb, text mining subsumes more shallow, simplistic analysis of text data that cannot for most purposes be considered machine learning. It mostly deals with representing large amounts of textual data in an insightful, simplifying way. Natural language processing, on the other hand, is able to grasp the semantic structure and context of textual data, learn relationships between words and terms, and create quantitative models of text to draw conclusions and create predictions.

Although Baker et al. (2016) demonstrated the power of text-based data, their method seemed rather cumbersome. Notably, the manual reading of large corpora to custom-identify the policy uncertainty words is not scalable to other macroeconomic questions. Azqueta-Gavaldón (2017a) has taken on this prominent issue and employed a technique known as Latent Dirichlet Allocation (LDA) to generate the different sources of economic policy uncertainty as 'topics' that resulted from this unsupervised machine learning algorithm. The author has shown that these topics matched the work of Baker et al. (2016) closely, and the resulting indices were almost identical. In Azqueta-Gavaldón (2017b), the author has subsequently applied the methodology for extracting financial and investment policy uncertainty in the UK and explored the narratives driving the cryptocurrency hype in Azqueta-Gavaldón (2020). Most recently, the ECB has worked on implementing the method of Baker et al. (2016) and Azqueta-Gavaldón (2017a) to analyse the driving forces behind policy uncertainty in the Eurozone in Azqueta-Gavaldón, Hirschbühl, Onorante, and Saiz (2020).

Outside of economic policy uncertainty, natural language processing tools can be found in the analysis of monetary policy communication, and in explaining inflation expectations formation. Schonhardt-Bailey and Bailey (2013) have examined two decades in "deliberations" by central banks – the mechanisms and reasoning on which they rest monetary policy decisions. Hansen, McMahon, and Prat (2017) attempt to examine whether such deliberative transparency can lead to substantially different monetary policy conclusions. In a related endeavour, Balke, Fulmer, and Zhang (2017) have analysed sentiment in FED Beige Books. Yet another example is the work by ter Ellen, Larsen, and Thorsrud (2019). The authors identified what they termed narrative monetary policy surprises by examining the importance of different topics (as inferred from an LSI – latent semantic indexing – model) in news media reporting relative to the periodically released monetary policy accounts of the Norwegian Central Bank. By measuring focus on a specific topic over time, the authors argue to have identified points in time where the central bank has provided a shock to the beliefs of market participants, provided that their focus was on a different area (topic) just before the press release of the central bank. The shocks identified in this process are shown to not correlate with the standard measures of monetary policy shocks, to consequently impact news media reporting, and to correlate positively with the 3-month interest rates and stock market outcomes.

A very different picture emerges beyond macroeconomics. In finance research, text mining and NLP have been used quite extensively already – especially in the context of capturing stock or another financial market sentiment, and in predicting their future evolution. There is a substantial body of literature on such studies, surveyed for example by Hagenau, Liebmann, and Neumann (2013), Khadjeh Nassirtoussi, Aghabozorgi, Ying Wah, and Ngo (2014) and Kumar and Ravi (2016). As their surveys show, a wide range of data sources, stretching from news articles, company reporting, social media text, speeches and corporate announcements have been used to predict financial market developments. Examining financial markets has one substantial advantage to the analysis of business cycles: the readily available and identifiable feedback loop between a prediction and the outcome in reality (e.g., of the stock price level in a few minutes after a news release). The information on market feedback allows the use of supervised machine learning techniques which can be trained on corpora employing past news releases and stock price realisations. Without a clear definition of business cycle stages and momentous feedback loop between text and macroeconomic variables, this will, arguably, never be possible in business cycle analysis. The tools I shall employ will, therefore, be limited to the realm of unsupervised machine learning. It is the learning of patterns from the textual data, which offers the only potential key to predicting business cycle developments.

Lastly, beyond literature in economics, there are several attempts at constructing text-driven early warning systems of crises in computer science research. As Tai, Olson, and Blessner (2016) review, most of it was restricted to supervised machine learning algorithms and focused on financial market variables such as the stock prices, exchange rates and gold price volatility. They were perhaps the first who used an unsupervised algorithm, the Latent Dirichlet Allocation, to construct topics from economic news reporting and consequently apply a classification tree in combination with cost adjusted prediction in a supervised ML model to finally predict crisis occurrence correctly in slightly more than half of the cases. In a different example, Choi and Varian (2012) focused on using Google Trends for nowcasting economic variables, for example, sales of companies or unemployment benefit claims and travelling patterns. Lastly, Rönnqvist and Sarlin (2017) use word vectorisation techniques, or word embeddings, to pre-learn semantics (context and word associations) from the news in an unsupervised fashion before predicting distress levels of banking institutions. The unsupervised technique of word embeddings shall take on a vital role in the analysis later on.

2.4 Natural Language Processing in Business Cycle Analysis

Language is a labyrinth of paths. You approach from one side and know your way about; you approach the same place from another side and no longer know your way about.

— Wittgenstein and Anscombe (1958, p. 82), *Philosophical Investigations*

Beyond the reasoning and evidence outlined in the previous three chapters on why text analysis should be an interesting avenue for economic research, there is one much simpler reason – the current macroeconomic models are not good enough. Brigden (2019) has recently shown, based on IMF macroeconomic projections, that the IMF economists were able to predict only four out of 469 recessions in 194 countries at least one year ahead. An, Jalles, and Loungani (2018) demonstrate that the private sector macroeconomic forecasts are equally 'good' at predicting – they were found to be nearly identical to the IMF forecasts. Parallel to this finding, there is an extensive body of literature dealing with the development of the so-called early-warning system (EWS) for economic crises. Babecký et al. (2013) reviews this work and offers a good overview of the quantitative indicators which are commonly believed to have business cycle predictive power. Housing prices and credit growth are found to be among the best values to watch out for. Another excellent example of a leading indicator is the treasury bond yield curve. The literature has provided pervasive evidence that this commonly known predictor of crises, more particularly the spread between the yield on short- and long-term treasury bonds, is significantly better at predicting recessions more than two quarters ahead than professional forecasters. The first to demonstrate this was Rudebusch and Williams (2009), who was subsequently confirmed by Lahiri, Monokroussos, and Zhao (2013). The effect was observed in spite of the fact that the predictive nature of the yield curve for future depressions has been known for some time. Why have the forecasters not successively learnt about the relationship and repeatedly neglected the measure as a good source of information on the future state of the macroeconomy? Why are the comprehensive, hardly constructed macroeconomic models used by professional forecasters not better predictors? These questions remain puzzling.

One answer could be that since the reasons for business cycle fluctuations – the nature of the shocks – change with each consecutive cycle, one can never build a model comprehensive enough to predict each innovation in the stage of the cycle, the mechanisms of action for a particular shock simply cannot be expected ex-ante. Textual data, on the other hand, could be different. The psychological information contained in them, and the real-time data availability, could provide clues for the future of the business cycle that quantitative measures, often known with a significant lag, could never provide. It is perhaps in this emotion and sentiment that one can find the unifying co-factor of (almost) *all* major business cycle turning points. How do we capture the boom and bust sentiment? As will become visible, the essence of this thesis will be just this NLP-based description of shifts in animal spirits.

The most closely related work is perhaps Tuckett and Nyman (2017) and Nyman et al. (2018). The authors of the latter employ Bank of England internal commentary on economic events and Reuters news reporting, among others, to demonstrate how relative shifts in the sentiment of these textual sources predicted the recent financial crisis. The authors focus on the relative roles of anxiety and enthusiasm in the news which they see as driving forces behind their *conviction narrative* paradigm outlined in Tuckett and Nikolic (2017). They argue that since market actors, like everyone else, live in a world of Knightian uncertainty, their desires to take risks are determined by their beliefs, convictions on the future state of the economy. These convictions are formed in a deliberative, structure-creating narrative process – the mind works to simulate the future, at each step working with one's knowledge and perception of the world, experiences and rules-of-thumb that are mixed with emotions to ultimately provide support to a specific action. This cognitive process essentially leads the agents to transform radical uncertainty into psychological certainty. The authors believe that these narrative convictions, should they embody animal spirits, are best achieved by focusing on sentiments that dictate approaching risk versus avoiding risk, which they see best represented in words suggestive of excitement versus that of anxiety[10]. Therefore, in

[10] A parallel that naturally comes to mind are the adjectives *dovish* and *hawkish* that often describe the communication on the need for expansionary and restrictive monetary policy respectively. These are also, arguably, expressions about the perceived future risk of a downturn versus an upturn by the central bank. In the spirit of Tuckett and Nikolic (2017), these could embody the emotional

Nyman et al. (2018), the authors create a dictionary of words expressing these two sentiments and measure the relative frequency of these words in the textual sources over time. It then turns out that a major shift in the *relative* proportional occurrence of excitement words versus anxiety words *preceded* the latest financial crisis.

Moreover, the authors find the economic news to be more consensual, as measured by entropy of the discovered latent topics in news reporting, before the crisis. Similarly to ter Ellen et al. (2019), these authors use Latent Semantic Analysis to identify the topics. Drawing on their insights from their conviction narrative theory in Tuckett and Nikolic (2017), they argue that the human mind creating the convictions can be in two different states at each moment – the divided and integrated state. Whereas in the latter we are receptive to new information and willing to shift our sentiments and thus update our conviction narrative, in the former, divided state, we disregard conflicting information and hinge on our previously formed beliefs. This unwillingness creates the seeds for the correction, and authors see the degree of this unwillingness embodied in the narrative consensus. Their theory implies that whenever economic beliefs align, the potential of a sudden, major correction in what the authors would call conviction narratives, driven by animal spirits, *increases*.

In parallel to this psychology-grounded approach to textual analysis in macroeconomics, there is the growing body of work on NLP-based *nowcasting* of macroeconomic variables. A particularly good example of this is the Norwegian Financial News Index (FNI) produced in a collaborative project between the Norwegian Business School in Oslo and the Retriever database of Norwegian newspapers. The methodological and theoretical know-how was developed by Thorsrud (2020) and Larsen and Thorsrud (2019b). The authors use a combination of LDA topic modelling, latent threshold dynamic-factor modelling, and sentiment analysis to derive an index that is competitive with other sources of macroeconomic nowcasting (incl. that of Norges Bank) and professional forecasts. Especially around business cycle turning points, the index in Thorsrud (2020) is shown to over-perform the GDP estimates of Norwegian statistical office which the author argues is an evidence of the informational (noise-reducing) value of news reporting. The constructed measure of present business cycle state draws on the notion of topic centrality – if a particular topic is central for news reporting in a time period, it is *assumed* to be important for the business cycle – and the simultaneous sentiment of the representative article of these central topics in each time period determines the value of the index. The same logic and modelling approach applies to the work of Larsen and Thorsrud (2019b). The Norwegian authors bring their work closer to the essence of narrative economics in their recent working paper in Larsen and Thorsrud (2019a). They bring much of the same logic and method to the analysis of US, Eurozone and Japanese economy, while additionally deriving an index measuring *virality* of narratives in economic news, and a measure of inter-connectedness and news spillovers between regions. Their work highlights that the set of narratives, as proxied by the topics, relevant for macroeconomic fluctuations *decreases* in troubled times, something that the authors see supported by the work of Nimark and Pitschner (2019). It is also in line with the finding by Nyman et al. (2018) that narrative consensus *increases* (entropy *decreases*) before a crisis.

These authors use the term narratives and LDA-topics interchangeably. They care to this critique by a discussion on the necessity of this approximation to make Shiller's concept of narratives in economics operational. Their work, however, aims to match topics to macroeconomic outcomes *by construction* as the connection between topic centrality and the macroeconomy is assumed and derived via VAR modelling. The connection thus does not arise completely exogenously; it is forced on the data. Therefore, albeit their work seems to address a similar research question as this thesis, their method, aims, as well as the identifying assumption as to what constitutes an economic narrative are quite different. In this thesis, the aim is to show that the computer can identify linguistic patterns which only happen to coincide with future macroeconomic outcomes in a *completely unsupervised* manner.

3 Text Corpora and Descriptive Statistics

The next step on the journey is the obtainment of a corpus of news reporting. *Dow Jones Factiva* was used to collect all corpora. The platform provides access to newspaper content from 200 different countries, in 28 languages, and claims to hold the most comprehensive news database in the world. Universities usually hold

states that dictate the overall conviction narratives of the central bank.

a user license – the use is, however, restricted to an estimated download allowance of under 1'000 articles per day which seriously limits the study. Comprehensive corpora of news reporting such as in Nyman et al. (2018) or Larsen and Thorsrud (2019a) are usually unavailable to students. A further goal of this thesis is thus to demonstrate that with a focused use of textual data, insightful research is still possible, yet definitely more difficult. The search query used to collect the corpus of news relevant for the research question was therefore of crucial importance, and the following lines shall demonstrate the reasoning with which this was done.

Business Cycle News Reporting (Corpus 1)

This corpus is amassed in order to learn how journalists, newspapers and those market actors reported about in them think about business cycles. What wording they use and how they use it. The articles should contain information about how people understand the economy, its macroeconomic principles, relationships between economic events and outcomes, and have business cycle relevance – the corpus should contain information on the model of the economy that the economic agents possess. How do we identify these articles?

> *We will refer to prosperity and revival as the positive phases of a cycle, to recession and depression as the negative phases.*
> — Schumpeter (1939, p. 156), *Business Cycles*

Based on Schumpeter (1939, p. 145-160), we can identify four stages of the business cycle: *prosperity*, *revival*, *recession*, *depression*. I shall refer to the first two stages as expansionary business cycle stages and the latter two as contractionary business cycle stages to refer to the direction of the overall trend in economic activity that they[11] refer to. Moreover, while writing about the cycle – which Schumpeter (1939) terms the *Secondary Wave* – he uses other expressions in roughly equivocal terms to the ones already mentioned: *recovery* (≈ revival), *expansion* (≈ revival and prosperity), *liquidation*[12] (≈ recession and depression) and *crisis* (≈ depression). Similar nomenclature is found in Rorty (1922, p. 78). From now on, I shall refer to these as business cycle *keywords*. These are central for gathering the corpus. Moreover, I attempted to introduce more commonly used terms that the media and journalists use to refer to a specific direction in the evolution of macroeconomic conditions. Consulting the thesaurus located at *dictionary.com* which serves as an aggregation of many scholarly accepted dictionaries of the English language, one finds the following synonyms of the term *economic expansion*: *boom*, *upswing*, *upturn* (thesaurus.com, n.d.-a). For the sake of balance, to have seven negative and seven positive words in total, their antonyms are identified as: (economic) *contraction*, *bust*, *downswing* and *downturn*[13]. In order to ensure the economic relevance of the news articles, they are also required to contain the words *economy* or *economic*. Both a noun and adjective form[14] of the word economy is used. This allows the query to find intuitive combinations such as *economic prosperity* or *economic downturn* and allows the identification of collocations such as *economy in a recession*. The search query and descriptive information on Corpus 1 can be found in Table 1.

Newspaper articles including these key terms, either in the headline or in the initial paragraph, are collected. I recognise that there is some ambiguity concerning the construction of this business cycle reporting query. While making the choice, I attempted to remain consistent, close to actual news reporting (reading of sample articles was performed), and choose an appropriate number of terms so as to be realistically able to collect the resulting number of articles given the usage limitations of *Factiva*.

The articles are collected from the following newspapers: *Reuters*, *The Wall Street Journal*, *The New York Times*, *The Washington Post*, *USA Today*, *The Financial Times*, *The Guardian* and *The Times (UK)*. These are chosen for their universal focus, wide readership and their broad coverage of economic issues. They are also generally accepted to produce professional content and are widely read by key market participants. As I recognise that some

[11] Figure 18 in Appendix A offers a graphical overview of the business cycle nomenclature.
[12] Using *liquidation* in the search query below did not result in almost any news articles over the 30 years period, so the word was discarded and not used in the final query.
[13] Alternative terms could have been included. The chosen ones seemed appropriate, given a reading of Merriam-Webster (n.d.-a) and Merriam-Webster (n.d.-b).
[14] The researcher could have preferred to include other grammatical forms of words such as adjectives, past participles, -ing forms – e.g., recovered, recovering – et cetera, but for the sake of not querying too many news articles and infringing upon *Factiva's* terms of conduct, I only used nouns in the query.

of these newspapers are U.S. based, and some are based in Europe, I use the American newspapers (*The Wall Street Journal*, *The New York Times*, *The Washington Post*, *USA Today*) to only extract news pertaining to the United States. The regional tag in *Factiva* is used for this purpose. Analogously, articles from the Europe-based newspapers (*The Financial Times*, *The Guardian* and *The Times (UK)*) are extracted, provided they pertain to a European Union or EFTA country (incl. the UK). *Reuters* is assumed as having reliable coverage of both regions. Both European and U.S. related news are collected here to have a sufficiently large corpus to learn on, as well as make the results more robust and universal by not restricting the pattern recognition to a specific single country.

Economic Expectations (Corpus 2)

This corpus of news reporting is collected in the belief that it offers textual information about markets' economic expectations at each point in time. It will be the corpus which shall ultimately be used to construct an index of business cycle sentiment, based on the patterns inferred from the corpus above, and used to construct an index of narrative consensus.

The thesis shall focus on economic expectations and index creation for the United States only. Therefore, only economic expectations news pertaining to the United States are used here. The regional tags provided by *Factiva* are used to query only U.S. related articles. The newspapers used here are *Reuters*, *Wall Street Journal*, *New York Times*, *Washington Post*, *USA Today*. The query is similar to the one above, consisting of terms pertaining to the economy in left parentheses and terms pertaining to expectations – 'forward-looking words' – on the right. In order to identify news articles with economic relevance, an expanded list of economic-relevance-indicating terms is used here: *economy*, *economies*, *economic*, *economics*, *economical*, *economist* and *economists*[15].

Subsequently, to define a sufficient number of diverse terms to identify *expectation* news, the following dictionary-based approach is used. When economic agents form expectations, they essentially predict the future in terms of forecasts. This sentence already gives us three important words: *expectation*, *prediction* and *forecast*. Using these core words, the synonym dictionary of English language – thesaurus.com (n.d.-b) – is consulted. Words judged closest synonyms by the thesaurus are extracted for both the noun and verb forms of the above three words. Afterwards, the following steps were applied:

1. All expressions in the following grammatical forms are included: noun (singular), noun (plural), verb, verb (third-person singular), -ing form.
2. Where the forms do not exist or are nonsensical, they are disregarded. Past tenses or passive speech forms are not included to strengthen the forward-looking nature of the search query.
3. *dictionary.com* or if needed other English dictionaries are checked for meanings of every single word. If the word is not overwhelmingly used for meanings related to outcomes in the future, i.e., forward-looking, and if it is not indicating a degree of uncertainty about the future, the word is disregarded. This step can be somewhat ambiguous – however, the choice of words is performed with caution.
4. The resulting words are checked individually, in combinations with the terms suggesting the presence of economic deliberations (the left parentheses), against Google search results and a sample of *Factiva* queried articles to make sure that they are used in overwhelmingly anticipatory contexts.

Please observe the resulting search query for Corpus 2 as well as related descriptive information in Table 1.

Figure 1 plots the article counts in each period. Corpus of business cycle news reporting peaks strongly around crises, albeit with some lag. This pattern should emphasise that the news reporting in Corpus 1 is related and representative of business cycles. The computer will use this corpus to infer linguistic patterns and ultimately

[15]The inclusion of the subject by the last two terms was deemed necessary to allow for the presence of some*one* forming the expectations. By including economists, and not financiers or households, I likely bias the sample of news towards that reporting on professional expectations. This choice should not weaken the core arguments of the thesis, especially if bearing in mind its explorative nature. For further analysis, it could be interesting to collect news reporting on economic expectations of businessmen or households specifically, particularly to examine the dimensions along which these differ.

learn the contractionary and expansionary lexica. Although the learning is performed on the *entire* corpus, the reader should note that robustness checks where only an early subset is used and several words are discarded from the lexicons will be performed in Section 5.4. As will be seen later on, most of the inferred words are remarkably generic, and there is no reason why they should not be identified, had we only used an earlier subset of news articles for learning. The second corpus article counts are rather stable over the 30-year period. This stability corresponds to the aim of obtaining a proxy for the prevalent expectations for *each* period. Corpus 2 directly underlies both the indices constructed later.

Table 1: News Corpora: Descriptive Information

Details About News Corpora		
Type of Information	Corpus 1 – Business Cycle News	Corpus 2 – Economic Expectations News
Search Query (terms in the headline or lead paragraph)	(economy OR economic) w/4[16] (prosperity OR revival OR recovery OR expansion OR boom OR upturn OR upswing OR recession OR depression OR crisis OR contraction OR bust OR downturn OR downswing)	(economy OR economies OR economic OR economics OR economical OR economist OR economists) w/4 (prediction OR predictions OR predict OR predicts OR predicting OR forecast OR forecasts OR forecasting OR expectation OR expectations OR expect OR expects OR expecting OR guess OR guesses OR guessing OR indicator OR indicators OR indicate OR indicates OR indicating OR prognosis OR prognoses OR prognosticate OR prognosticates OR prognosticating OR prophecy OR prophecies OR prophesy OR prophesies OR prophesying OR anticipation OR anticipations OR anticipate OR anticipates OR anticipating OR envision OR envisions OR envisioning OR foresight OR foresights OR foresee OR foresees OR foreseeing OR estimate OR estimates OR estimating OR outlook OR outlooks OR projection OR projections OR augur OR augurs OR auguring OR foretelling OR foretell OR foretells OR portent OR portents OR portend OR portends OR portending OR prospect OR prospects OR await OR awaits OR awaiting)
Description	This corpus encompasses news reporting relating to business cycles specifically. It is designed to contain articles that have mentioned several business cycle keywords prominently in the text.	This corpus encompasses news reporting relating to economic expectations, predictions and forecasts. It is by design overwhelmingly forward-looking.
Time Horizon	January 1990 to April 2020	September 1987 to April 2020
Newspapers	*Reuters*, *The Wall Street Journal*, *The New York Times*, *The Washington Post*, *USA Today*, *The Financial Times*, *The Guardian* and *The Times (UK)*	*Reuters*, *The Wall Street Journal*, *The New York Times*, *The Washington Post*, *USA Today*
Regional Coverage	USA and European Union Countries, incl. the UK and EFTA	United States of America
Source	*Factiva* (n.d.)	*Factiva* (n.d.)
Counts	Corpus 1 – Business Cycle News	Corpus 2 – Economic Expectations News
Number of Articles	61'675	31'259
Number of Sentences	910'783	499'588
Number of Words	31'502'193	14'444'433
Number of Words (after trimming)[17]	15'475'881	7'127'147
Number of Unique Unigrams	278'470	146'639

[16]The 'w/4' refers to the functionality of *Factiva* to search for articles where the words defined in the parentheses to the left and to the right are at most a four words distance from each other in the text.

[17]Trimming refers to the removal of stop-words, special characters, nonsensical strings, etc.

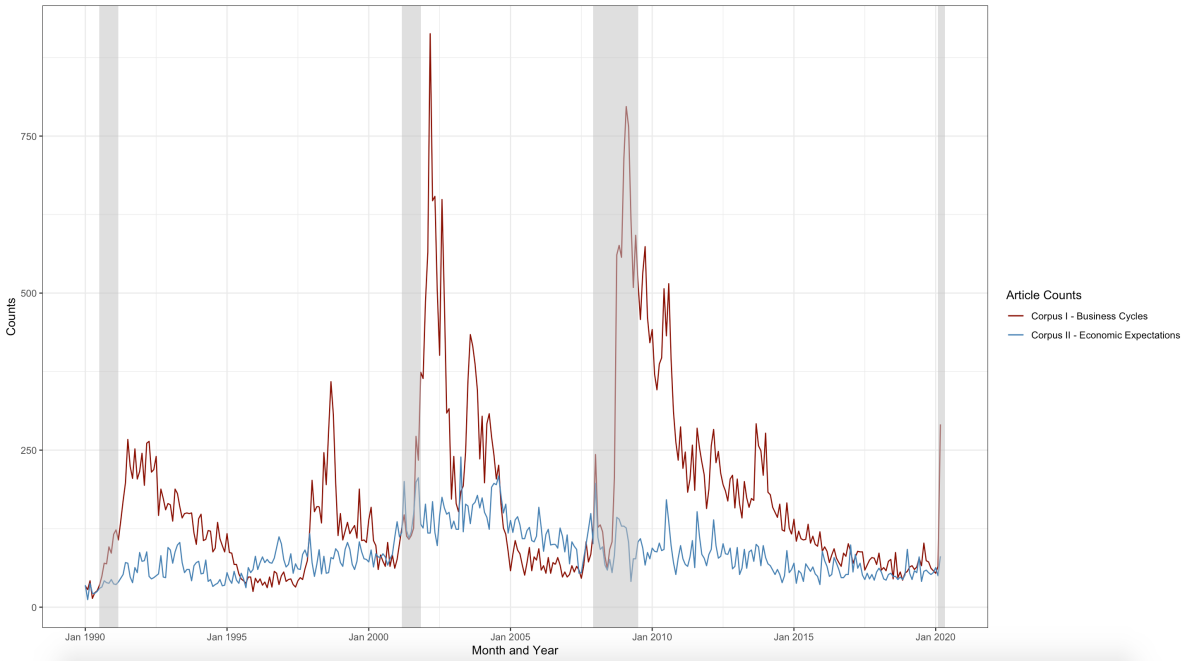


Figure 1: Monthly Article Counts: Corpus 1 and 2. Plotted Starting January 1990. Grey Areas are NBER-Defined U.S. Recessions (National Bureau of Economic Research, n.d.).

4 Methodology

The motivation for this chapter is two-fold. Firstly, it is necessary to accurately present how I arrived at the results. Secondly, and perhaps more importantly, I want to present techniques that are seldom used in the field of macroeconomics but could prove fruitful, and so encourage their use. Figure 2 should aid the reader in structuring this chapter. The first two steps of text analysis, pre-pre-processing and information retrieval, have already been performed above. Section 4.1 will commence by discussing relevant special properties of textual data to motivate the need for pre-processing, text mining and NLP in this thesis. Section 4.2 then dives into pre-processing and briefly explains how the data was amended and handled before the algorithmic analysis later on. Discussion of text mining is omitted – however, note that text mining tools were used along the way, whether in terms of simple word clouds, basic sentiment analysis, or in the course of explorative analysis of the gathered corpora. Section 4.3 will progressively develop the techniques needed to understand the construction of the two indices: Relative Business Cycle Sentiment and Narrative Consensus. It features Sections 4.3.1 and 4.3.2 that introduce matrix factorisation, Latent Dirichlet Allocation, topic modelling and word vectors *in general*. Crucially, Section 4.4 then explains the methodological application of these techniques and algorithms in the index construction. The reader already well-experienced in textual analysis, SVD, LDA, and word vectorisation techniques may wish to skip parts of the discussion in 4.1–4.3. Lastly, Section 4.5 introduces the evaluation of the indices, the comparative time series collected, and the econometric framework used. Along the way, further intuition for the text processing steps is developed and noted in Appendices B–F. Finally, the explanations of the econometric framework used are communicated in Appendix H and I.

There is a substantial amount of degrees of freedom when working with textual data. Many decisions could have been made differently and are non-trivial to justify or evaluate. The econometric evaluation and the raw output from text mining and NLP generate a feedback loop between the researcher, the algorithm, and the results. To emphasise this point, Figure 2 features two iterating arrows below text mining and natural language processing as well as two 'feedback' arrows between TM and NLP. The method here presented should, therefore, be interpreted as a current best guess on a useful approach that is worthwhile to explore further. Programming language R was chosen for this thesis. The interested reader is welcome to explore Kwartler (2017), Silge and Robinson (2017), Manning and Schütze (1999) or Hansen (2019) as further sources for studying computational textual analysis.

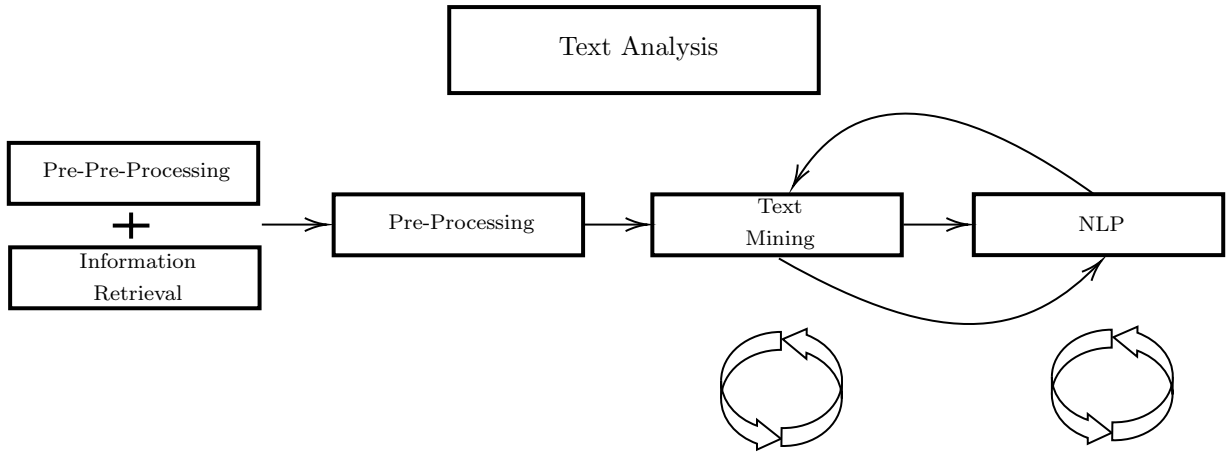


Figure 2: Segments of Textual Analysis and Their Sequence

4.1 What Makes Textual Data Special?

Human language is a wide and complex network of words, meanings, and (often informal) rules that developed over millennia. Teaching the computer to understand it involves both linguistic and computational difficulties. On the linguistic end of issues, there are two with particular relevance for the research question – vocabulary mismatch and ambiguity of terms and words. Early on, Furnas, Landauer, Gomez, and Dumais (1987) have shown that people can use a great variety of words to describe essentially the same thing. In their study, the same term was used with less than 20% probability to describe precisely the same concept. This finding applies to academic concepts as well as to commonly used words. If different people use different vocabulary to describe the same concepts, how can computer identify an economic narrative or understand what the nature of a belief is? The computer needs a way to categorise words into sensible baskets that we can meaningfully interpret, and use for further analysis. This is why latent structure modelling is of central importance in unsupervised machine learning and will be discussed in Section 4.3.1.

Whereas the first issue corresponded to the problem of different words having the same meaning, language ambiguity refers to one word having many meanings. The ambiguity can manifest itself in lexical, semantic or syntactic ways. Where the first, lexical, refers to the many dictionary definitions of meaning, the second, semantic, refers to situations where subject to its context, a term might be understood in a variety of ways. The last emerges where the ambiguity results not from different possible contexts, but from different possible sentence structures. In conclusion, inferring the meaning of words is no simple task for a computer, and it needs a metric to define and differentiate the meanings of words. Inferring meaning from linguistic patterns is one of the motivations for the emergence of context-based techniques and the ultimate reason for the development of vector space modelling (word embeddings) that will be discussed in 4.3.2. The problem is often referred to as word-sense induction. For our purposes, the computer should infer how the meaning of business cycle keywords relates to the meanings of other words so that we ultimately can find an expansionary and contractionary vocabulary.

Computational challenges form the second group of issues and can be subsumed with the notions of multidimensionality and sparsity of textual data. The former refers to the broadness of language vocabulary[18] as when each unique word needs to be represented by a number. The latter subsumes the fact that only a tiny subset of all possible vocabulary words tend to occur in a specific text, which results in a matrix representation of corpora where the great majority of entries is zero. Therefore, the decision on how to represent data and what data to discard is crucial. We want to successively reduce the number of entities, such as words, that we analyse. This step is where data pre-processing becomes relevant. We have to define clever ways to subset, and clean the

[18]If we imagine having 50'000 news articles, consisting of 300 words each, we would already end up with 15'000'000 tokens that have to be represented mathematically. According to Merriam-Webster (n.d.-c), the English language boasts around 470'000 words. On top of that come the different grammatical forms and various special characters.

corpus to both make the computations simpler, and the results more accurate. The critical aim here is to remove information that does not allow us to discriminate between entities (e.g., news articles) in the corpora and the information which is either unnecessary or introduces unwanted noise into the prospective conclusions.

4.2 Pre-Processing Textual Data

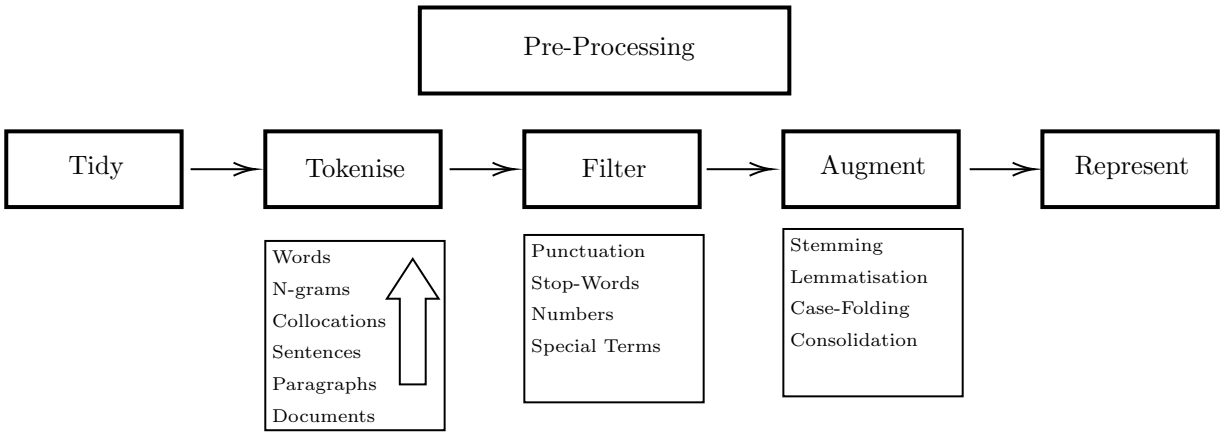


Figure 3: Steps Involved in Common Pre-Processing of Textual Data

The pre-processing steps follow the chart in Figure 3 and will be summarised below. Further details can be found in Appendix B.1. As highlighted by the previous chapter, two corpora of news articles were created. One pertaining to reporting about business cycle stages and one pertaining to economic expectations. The articles provided by *Factiva* are loaded into R, and all information is retained – most importantly, the time stamp of each article, its headline, newspaper source, regional tags and text. Headings and respective bodies of all articles are concocted for the corpus on economic expectations, for the corpus on business cycles this is not done for reasons elaborated later on. The text of each article is subsequently split into paragraphs, sentences, bigrams and words (unigrams). The latter are the main blocks of data used for the later analysis. Stop-words are removed in the process of tokenisation of sentences into bigrams and words. The stop-word lexicon called *onix*[19] was used. The lexicon was checked not to include any economic terms, or terms that could significantly impact the conclusions of the analysis later on. Stop-words specifically related to *Factiva's* database, such as websites or e-mail markers, nonsensical words, etc. that were found in the corpus are removed as well. Numbers and counts are also removed to the best of my ability. The words are *neither* stemmed nor lemmatised as this was found to result in loss of valuable information.

The resulting data on unigrams is stored in a document-term matrix (DTM) with raw token frequencies for the analysis with Latent Dirichlet Allocation. Word embeddings are created based on information from a term-term-co-occurrence matrix (TCM) based on unigrams and bigrams. Example of a DTM can be found in Figure 4 and of a TCM in Figure 7. The former is commonly referred to as *bag-of-words* approach, or *bag-of-words* model and the data underlies the Narrative Consensus Index. The latter is a *context-based* approach as the matrix captures the context of the tokens, and can be thus used in the analytical evaluation of the *semantics* of a language. This matrix underlies what will become the Relative Sentiment Index. Once the desired textual representation has been generated, the researcher can then unleash the variety of possible natural language processing tools. The subsequent chapter focuses on natural language processing algorithms. For further detail on pre-processing, kindly refer to Appendix B, with data matrices being elaborated in B.2.

[19]The list of words removed can be viewed at https://www.lextek.com/manuals/onix/stopwords1.html. Examples of stop-words would be *the*, *a*, *that*, *and so on*.

4.3 Natural Language Processing

Sections 4.3.1 and 4.3.2 will progressively develop the reasoning needed to understand the construction of the two indices: Relative Business Cycle Sentiment and Narrative Consensus. The subsequent chapters introduce matrix factorisation, Latent Dirichlet Allocation, topic modelling and word vectors *in general*. Topic models will help to structure economic news into categories, and the relative focus on these categories through time will later be used as a proxy for narrative consensus, or broadness of economic dialogue. Afterwards, word vectorisation is introduced, where it shall become clear how the prevailing business cycle sentiment of each stage of the cycle can be captured in terms of a representative, narrative vector.

4.3.1 Discovering Hidden Structures and Variables: LSI and Topic Modelling

The group of NLP techniques discussed here has been the most commonly used one in economic literature to date. What binds them together is their focus on hidden, or *latent*, structures in data. What are these latent structures, and why should we be interested in them? Surprisingly, the answers to these crucial question are diverse and somewhat unclear in the literature on machine learning. It is far from obvious what a 'latent structure' is supposed to represent. However, their objective is clear. If one is faced with a large amount of unstructured data (as with newspaper articles in this thesis), the researcher wants to develop techniques to search and organise, and ultimately to establish connections between news articles and cluster them in insightful ways. Virtually all unsupervised learning algorithms in NLP are based on this underlying motivation. If one compresses the information in a matrix based on textual data, such as the Document Term Matrix, in the right way, interesting patterns emerge. The word vector representation, and vector space modelling in general that will be introduced in the next section also hugely builds on this logic. The machine 'learning' happens right here, in an appropriate compression of the information in the data matrices.

Commonly, the document-term matrix is compressed via a procedure known as Singular Value Decomposition which leads to the Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) model developed by Dumais, Furnas, Landauer, Deerwester, and Harshman (1988). Appendix C and Dumais (2004) feature an accessible non-mathematical overview of the technique. Figure 4 highlights the process. Of the economic literature reviewed in 2.3 and 2.4, ter Ellen et al. (2019) and Nyman et al. (2018) have used the algorithm. This compression is, however, only the first step in a broader analysis. One still needs to explicitly find 'clusters', such as clusters of similar documents, given by these latent components. These clusters can be determined by some standard measure of distance, such as cosine distances. Alternatively, the vector space given by SVD can be used together with conventional clustering algorithms such as k-means or is otherwise transformed for the researcher's purposes. Nyman et al. (2018) have applied x-means clustering, and ter Ellen et al. (2019) have rotated the resulting singular vectors in a way as to be able to interpret them as inflation narrative in a custom method they developed.

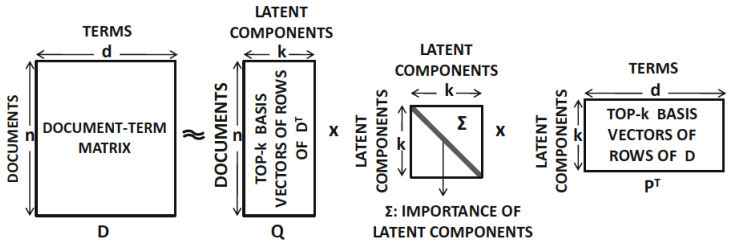


Figure 4: SVD as visualised by Aggarwal (2018, p. 37).

Many other NLP tools emerged after the seminal contribution by Dumais et al. (1988). Most notably, *Non-Negative Matrix Factorisation* (NMF), *Probabilistic Latent Semantic Analysis* (pLSA) and ultimately, the *Latent Dirichlet Allocation* (LDA), which I shall be using, was developed. These latter two models represent a category

of machine learning models called probabilistic generative models. They do not use the SVD to infer the latent variables from the data – instead, they use Bayesian statistics and inference. Based on pre-specified *prior* beliefs, the algorithm generates (conditional) posterior distributions of these hidden variables of interest given the observed data[20]. While the term probabilistic being obvious, these models are called generative precisely because of this Bayesian derivational aspect; one assumes that the tokens in documents and the latent components are generated by an a priori assumed distribution. It is as if the documents and categories (topics) were coming from a latent distribution which existed before the data was observed. For example, in the case of LDA, the distribution over the hidden variables is assumed to be generated by a prior Dirichlet distribution. Both pLSA and LDA decompose the observed reality into a mixture of distributions that describe the latent classes, referred to as mixture models, or latent class models. Only such probabilistic models of text are referred to as *topic models* in the NLP literature.

The LDA was the first that allowed mixture modelling on both the document and the word level with comparably little computational effort, which rendered it widely used. Particularly, it assumes a likelihood function that generates each word w_n in a document $\mathbf{w}$ where n, d are used to index specific words and documents, N_d, M are the upper indices for the last word and last document in the corpus respectively and α, β are hyperparameters of the Dirichlet distribution. The likelihood function is noted to have the form:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n} p(z_n|\theta) p(w_n|z_n,\beta) \right) d\theta \tag{1}$$

One can understand $p(\theta|\alpha)$ as a probability distribution over topics in a document $\mathbf{w}$. Note the boldness of w; it should emphasise the document consisting of a collection of tokens (bag-of-words). θ is a multinomial distribution of topics in a document. The two multiplied probabilities in the sum over topics z of a word n in equation (1) determine the specific, observed incidence probability of a single token in the document, given the token was coming from the topic z_n. Summed over z_n, this gives the probability of observing the single token. Note that each word is, in the end, assumed to have come from two distributions, indirectly from a Dirichlet distribution (hyperparameter: α) that pins down the multinomial θ from which the realisations of the categorical z are coming, and directly from another Dirichlet (hyperparameter: β) that gives the distribution of words in each topic.

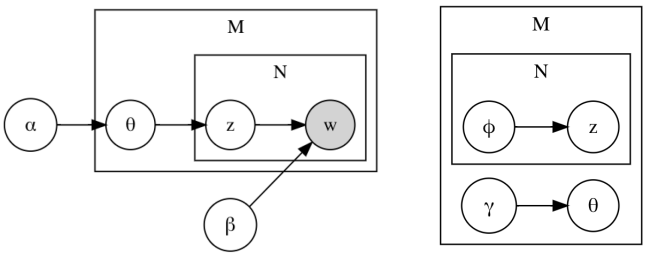


Figure 5: Inference in the LDA model. Source: Own Diagram. Inspired by Blei et al. (2003).

LDA maximises the likelihood of generating the observed documents $d = 1, ...M$ given the likelihood function of generating a document in (1):

$$l(\alpha,\beta) = \sum_{d=1}^{M} log\left(\mathbf{w}_d|\alpha,\beta\right) \tag{2}$$

The inference is performed by consecutively sampling, and ultimately by variational Bayes. The sampling is presented below so that the reader gets further intuition into how LDA generates the observed documents from the assumed latent structure.

[20] The reader interested in the basics of Bayesian statistics is encouraged to consult Koop (2003) or Gelman et al. (2013).

1. Draw $\theta_d \sim \text{Dirichlet}(\alpha)$. This gives us one realisation from the distribution over latent variables (topics) k in a specific document d.
2. Draw $\phi_k \sim \text{Dirichlet}(\beta)$ where the realisation of ϕ_k gives us distribution of words in a specific topic k.
3. Then for every word in N, iterate two steps:
   (a) The drawn θ_d for the document is used to choose a topic for a word n by drawing $z_{d,n} \sim \text{Multinomial}(\theta_d)$. Note how it is assumed that a word is generated from a single topic k.
   (b) Conditional on the drawn $z_{d,n}$ and ϕ_k, a word is drawn from $p(w_{d,n}|z_{d,n}, \beta)$, or equivalently $w_{d,n} \sim \text{Multinomial}(\phi_{z[d,n]})$

Sampling in the LDA Model

Two key variables are ultimately inferred in the model. The first one approximates the topic distribution of each document, and the second approximates the probability distribution over words in each topic. The reader can think of these as posterior counterparts of the prior distributions given by the α and β hyperparameters, albeit note that the reality is considerably more complex. The original paper notes that the γ distribution, which is used to denote the posterior distributions of topic incidence in the observed universe of documents is given by $\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$ where i is a particular component of the distribution and ϕ_{ni} describes another parameter of the variational distribution and has to be inferred via a convergence algorithm. In simplistic terms, the reader can think of the γ-distribution as the probability $p(z|\mathbf{w}_d)$ for each d. This distribution plays a key role in Section 4.4.3 and is used to create the measure of narrative consensus via its entropy.

Secondly, it is the posterior β distribution which is found having the following kernel: $\beta_{i,j} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}$ where d, n, i and j are the indices for documents, specific words (tokens) in each document, the elements of the approximated variational parameter ϕ, i.e., indices of specific topics, and lastly, j stands for each unique word in the overall vocabulary of the universe of documents. It is a distribution describing a topic by the probabilities of words representing it that could be denoted to consist of $p(w^j|z)$ for each j. Appendix K features Figure 31 that plots some of the words representative of several topics – those with highest β-value for each z given the corpus on economic expectations.

Both variational Bayes methods and Gibbs sampling techniques have been applied in the literature to derive these key variables. In this thesis, the inference is made exactly as in the original paper of Blei et al. (2003), i.e., by variational inference. The outline here follows and is inspired by Blei et al. (2003). The interested reader can refer to Appendix D that provides further intuition, or to the original paper. For a mathematical exposition of the concepts needed to understand these machine learning models, a good starter is Strang (2019). To find out more, Aggarwal (2018) reviews many of these machine learning models with solid mathematical treatment. In this study, the R package *topicmodels* by Grün and Hornik (2011) implementing LDA is used.

4.3.2 Understanding Context: Word Embeddings and GloVe

The previous section has evolved around a type of NLP models that all share a particular, important assumption. The latent variable models all assumed the context of individual tokens not to matter, something that was referred to as a *bag-of-words* model. Remember that the input to all algorithms disregarded the token *order* altogether. Arguably, this could be a hugely important aspect of economic storytelling. Think about the task at hand – the thesis wants to teach the computer how newspapers report, and thus think about business cycles.

Let us consider the methodological approach of this thesis. At each point in time, the economy finds itself in a specific business cycle stage, somewhere along the business cycle curve. In the most simple mutually exclusive nomenclature, at every point in time, the economy finds itself either in the expansionary or in the contractionary

phase as defined by National Bureau of Economic Research (n.d.)[21]. On the more objective end, the quantitative economic data such as GDP growth or unemployment rate give us a picture of the economy that provide clear hints as to in which of these stages we currently are. Beyond the realm of quantitative data, there are the beliefs and stories we tell ourselves regarding the current state of the macroeconomy. In newspaper reporting, these economic beliefs are implicitly contained in words, their order, context and tone – in the semantics, or meaning, of the text. Therefore, to understand the current state of the beliefs about the business cycle from news reporting, we need to understand the semantics of the written language. The questions then offer themselves: How can we approximate semantic understanding? What feature of text captures its meaning? The English linguist John Rupert Firth provides a clue:

> *You shall know a word by the company it keeps.*
>
> — Firth (1957, p. 179), *A Synopsis of Linguistic Theory*

What Firth writes sounds simple, yet it has important implications. If the computer is to understand the meanings of words, it needs to model their context. Therefore, if we are ever to represent an economic narrative from a text, or track its evolution, the key could lie in statistically describing the properties of contexts of words, and in comparing the properties of different contexts to one another. In computer linguistics and statistical semantics, this is commonly referred to as the *distributional hypothesis*. The recently introduced approaches to the word vector representation of text build exactly on this logic. These are unsupervised algorithms whose key object of learning is the context of tokens. If the computer is able to sensibly predict contexts of words, it is reasonable to claim that the computer somewhat understood the semantics of a word or any textual unit it learnt to predict[22].

If you were asked the question, "Do you believe that the economy will expand or contract next period?", you will most probably respond with different vocabulary depending on whether you believe the former, or the latter – especially if you elaborate on your reasons why you believe so. The key for the computer to understand the nature of your beliefs, then, lies in recognising the linguistic differences in your potential answers. By the distributional hypothesis, these differences can be described by relative probabilities of term co-occurrences. For example, if you believe that the economy will contract the next period, your answer to the question above is much more probable to include the words *worsen*, *negative* or even *crisis* than words such as *prosper*, *blossom* or *football*. Evidently, if the journalists writing news articles about economic expectations have different beliefs, they will tend to use different vocabulary. However, we still need to set a target term, the context of which is of our interest. For the research question at hand, this is a set of *unambiguous* words that are regularly used to describe a specific state of the macroeconomy. These business cycle keywords were already introduced in Chapter 3 (cf. the *Factiva* search query used to collect business cycle news) with the most prominent examples being *expansion* and *contraction*. They are mutually exclusive by their economic definitions and, on themselves, have crystal-clear sentimental value. Suppose we were to predict the sentences "The trade war between U.S. and China will lead to economic ___" and the sentence "The blossoming trade between U.S. and China will lead to economic ___". In that case, the conditional probability of the word *expansion* coming after economic, given the entire context – all other words in the sentence, is higher in the latter sentence, and vice versa for *contraction*. What is the main difference between these two sentences? It is the word *war* in the first, and the word *blossoming* in the second sentence, that make up the difference in contexts and thus meaning, between expansion and contraction. It is then straightforward to argue that if the word *war* occurs relatively more often in economic expectation news – especially more often than the word *blossoming*, the beliefs of the journalists writing the articles are more representative of an economy in contraction. Just as we determined the semantic association of war and contraction, and of blossoming and expansion, we can identify an entire lexicon of words that are semantically similar to key business cycle words.

By learning on the contexts of different business cycle keywords in economic news reporting, we are not only capable of describing differences in linguistic patterns along the business cycle, and thus the belief sentiment as such, but we can focus our analysis on the vocabulary used by *specifically* these newspapers, by their journalists,

[21] Consult Figure 18 in Appendix A for an overview of the business cycle and the keywords the thesis is using to describe it.

[22] Example: Nowadays, Google's e-mail offers sentence completion suggestions while users are typing. This is exactly the sort of language understanding a computer needs to possess to understand the meaning of words.

in this macroeconomic business cycle context, and in a specific time period. All this would not be possible with a generic sentiment lexicon, created to superficially fit a broad, different array of texts. Much more likely would the analysis with a generic sentiment lexicon become noisier as words of minor or unclear macroeconomic relevance could skew the sentiment score[23], and the sentiment of some words could even be wrongly assumed[24]. Increasingly, it is becoming recognised that various domains and communities use not only different vocabulary, but also that the same terms might be of different sentiments and meanings in different communities and domains. The recent work in machine learning literature underlines this point (e.g., Deng, Sinha, & Zhao, 2017; W. L. Hamilton, Clark, Leskovec, & Jurafsky, 2016; Huang, Niu, & Shi, 2014)[25].

This thesis uses the context-learning word vectorisation algorithm of Pennington et al. (2014) called GloVe, or Global Vectors for Word Representation. The algorithm takes a term-term co-occurrence matrix, extracts information on relative probabilities of words occurring in different contexts, and reduces this information via a low-rank approximation. An interesting property of GloVe is that it uses *global* corpus statistics of word co-occurrences while learning, which stands in contrast to similar procedures that merely learn on local context-window statistics. Note how this corresponds to the second-mentioned proposition of Shiller (2019) in Section 2.2. As an introduction, observe Figure 6 taken from the original paper.

The last row of Figure 6 operationalises Firth's idea of where meaning comes from. Given the context word $k =$ solid, the probability of observing word $i = ice$ is much higher than that of $j = steam$, and thus the relative probability is a relatively large number. The opposite is true for *gas*. For $k =$ water and $k =$ fashion the ratio is uninformative, in the sense that either both or neither are semantically (or syntactically) related to *ice* and *steam*. This information about relative co-occurrence probabilities of *all* words in the corpus is later encoded in a multidimensional vector space where the vector embedding for *ice* can be found relatively close to *solid*, and *steam* will turn out to be relatively close to *gas*. Two possibly unknown terms were introduced here – embeddings and vector closeness. Mathematically, a vector embedding is a structure-preserving mapping of a discrete variable (a token) to a vector of continuous numbers. Secondly, when speaking of closeness, I refer to geometric closeness as defined by the angle defined by these vectors. It is this angle[26] which defines the semantic and syntactic similarity between two tokens in a corpus. In the following paragraphs, I will use the words *token*, *word* and *term* interchangeably.

Table 1: Co-occurrence probabilities for target words *ice* and *steam* with selected context words from a 6 billion token corpus. Only in the ratio does noise from non-discriminative words like *water* and *fashion* cancel out, so that large values (much greater than 1) correlate well with properties specific to ice, and small values (much less than 1) correlate well with properties specific of steam.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Figure 6: Meaning in Term-Term Co-Occurrence Matrices: GloVe Learning. Table taken from Pennington et al. (2014).

In this study, the R implementation by Selivanov, Bickel, and Wang (2020) of GloVe shall be used. Based on the original paper, word embeddings are approximated in the following steps (cf. Pennington et al., 2014):

[23]Example: Word *decrease* would be given a negative score in a generic lexicon, but can by no means mostly imply something economically negative. For example, the article could write about a decrease in inflation. Does it suggest a rather contractionary or expansionary sentiment? No simple, if any, answer is possible.

[24]Example: Words *fledgeling* or *feeble* are negative on its own but often co-occur with the word *recovery*, and thus by Firth's distributional hypothesis they bear meaning similar to that of recovery.

[25]This thesis will not evaluate the success of the constructed text-based indices in comparison with common sentiment lexicons. The reader is referred to Appendix J, Figure 29, for visual comparison with standard sentiment scores.

[26]The magnitudes (or equivalently, the L2 norms) of embeddings are in general highly variable, and therefore comparing words by standard Euclidian distance makes no straightforward sense.

1. Researcher specifies hyperparameters: The dimensionality of the to-be-estimated embeddings, the context window to be evaluated, the n-degree of N-grams to be considered, and the upper limit of the weighing function. See Table 2 for choices.

2. Generate a term-term-co-occurrence matrix $\mathbf{X}$ where each element $X_{i,j}$ records how often a term i appears in context window around another term j. See Figure 7 for an example.

3. Extract 'meaning' by means of relative probabilities of token occurrences in the context of other tokens: $\frac{P_{ik}}{P_{jk}} = \frac{P(k|i)}{P(k|j)}$, where $\frac{P(k|i)}{P(k|j)}$ is equivalent to the last row of Figure 6, $i, j, k \in \mathcal{V}$ are words from the vocabulary and we know that $P_{ik} = \frac{X_{ik}}{X_i}$ where X_{ik} and X_i are from the term-term-co-occurrence matrix.

4. The contribution of Pennington et al. (2014) is to show how to derive the word embeddings $\mathbf{w}_i, \tilde{\mathbf{w}}_k$ [a b] from the information encoded in $\frac{P_{ik}}{P_{jk}}$. They postulate a function F so that $F(\mathbf{w}_i, \mathbf{w}_j, \tilde{\mathbf{w}}_k) = \frac{P_{ik}}{P_{jk}}$ and derive that under certain assumptions $F(\mathbf{w}_i^T \tilde{\mathbf{w}}_k) = P_{ik}$.

5. The paper shows that one can assume F to be an exponential function and then reduce the information about P_{ik} from the term-term-co-occurrence matrix $\mathbf{X}$ via a low-rank approximation to $\mathbf{w}_i$ and $\mathbf{w}_k$, while attempting to fulfil the following constraint: $\mathbf{w}_i^T \tilde{\mathbf{w}}_k + b_i + \tilde{b}_j = log(X_{ij})$ where the tilde is for the context words and b_i and $\tilde{b}_j$ are bias terms. Note how the word vectors are forming a dot product, and thus result from a factorisation of the logarithm of matrix $\mathbf{X}$. Reader interested in properties of matrix factorisation is referred to Appendix D.

6. To optimise the embeddings (so that their error in representing the term co-occurrence matrix is minimal), the following cost function is minimised via stochastic gradient descent: $J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\mathbf{w}_i^T \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - logX_{ij} \right)^2$. $\mathbf{w}_i, \mathbf{w}_j$ are the embedded word vectors, b are biases, V is the size of the vocabulary set $\mathcal{V}$, and the function f is a weighting function to make sure that words co-occurring most often or only a few times do not get over- or under-weighted respectively. Less weight is given to more distant terms.

[a] Note that $\mathbf{w}_i$ and $\tilde{\mathbf{w}}_k$ are now vectors of numbers representing a specific *single* token with dimensionality specified in Table 2. In the previous section, $\mathbf{w}$ was used to denote a *vector of words* – an entire document – and w_n was a single word.

[b] Note that each token is both a target and a context token – depending on the perspective. Therefore, each token is represented both in terms of $\mathbf{w}_i$ and $\tilde{\mathbf{w}}_k$. These two vectors are commonly either summed or averaged at the last step of the procedure to deduce a final, singular, embedding for each token.

$X =$

	I	*like*	*enjoy*	*deep*	*learning*	*NLP*	*flying*	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

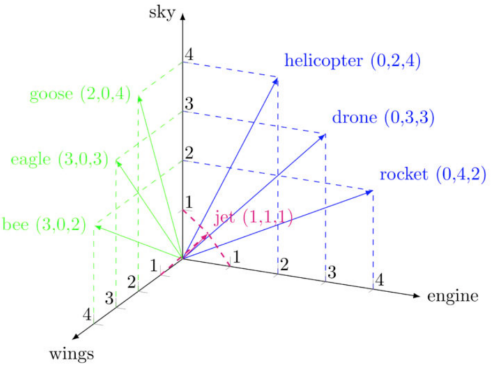


Figure 7: **Left**: Example of a term-term-co-occurrence matrix. Given by the sentences "I enjoy flying. I like NLP. I like deep learning". Taken from Chaubard et al. (n.d.). **Right**: Example of a generic, non-embedded word vector space that is defined by the relationship between three context tokens and seven target tokens. Coordinates are given by counts of target words in a window around the context words on the axes. Taken from Desagulier (2018).

This procedure completes the transition from the term-term co-occurrence matrix through relative co-occurrence probabilities to embeddings. The term-term co-occurrence matrix $\mathbf{X}$ that is of remarkable dimensionality[27] gets reduced to a 100 dimension vector in an $\mathbb{R}^{100}$ space where a 100×1 vector represents each word. Intuitively, if one forces the matrix $\mathbf{X}$, which has dimensions of the length of vocabulary to a space of 100 dimensions while minimising the loss of information, the most important relationships between words must be preserved. Patterns get discovered, similar to those which emerge from the SVD factorisation in LSI – in the form of a latent structure[28]. This latent structure encodes the information about relative word-co-occurrence, and could be imagined as a distribution over contextual words – where the context is represented not in its entirety, but approximated via this latent low-dimensionality structure. Furthermore, by focusing on *relative* co-occurrence *probabilities* and overweighing lower co-occurrences, the algorithm takes implicitly account of the proposition made by Shiller (2019) that narratives could comprise only a small percentage of talk, as outlined in Section 2.2.

In this final vector space, both the direction and magnitude of the embeddings bear certain information about the mutual relationships of the words to one another. As is widely observed in the literature on word vector space models (e.g., Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington et al., 2014), the estimated vector spaces feature meaningful *linear* relationships between tokens. Most prominently, algebraic operations on the vectors allow constructing paraphrases, synonyms and analogies. There is still significant discussion in the computer science literature as to how these relationships emerge, and why. The emergence of these spatial relationships is discussed in Levy and Goldberg (2014), Arora, Li, Liang, Ma, and Risteski (2016), Gittens, Achlioptas, and Mahoney (2017), Allen and Hospedales (2019), Allen, Balazevic, and Hospedales (2019) and Ethayarajh, Duvenaud, and Hirst (2019).

The existence of semantic linear substructures will be used to construct two vectors capturing the expansionary and contractionary narrative. For this purpose, it is important to understand how vector summation and subtraction boils down to *literally* meaningful operations on word embeddings. Let us start with the summation of word embeddings. It has been shown that $\mathbf{w}_{\text{man}} + \mathbf{w}_{\text{royal}} \approx \mathbf{w}_{\text{king}}$ (e.g., Ethayarajh et al., 2019). How can this be? Gittens et al. (2017) cleverly call this relationship a *paraphrase*. The intuition is that any set of words, say a set of context words C where $C = \{man, royal\}$ could be taken, a perfect paraphrase would be a token C_x such that

$$P(W|C) = P(W|C_{\text{x}}) \tag{3}$$

where W can be any word in the full vocabulary $\mathcal{V} \setminus \{C, C_x\}$. The intuitive explanation is that if the meaning of each token is given by a probability distribution over words that co-occur in its vicinity, a perfect paraphrase to a set of words would be a word that has the most similar probability distribution over the contexts as the set of words to be paraphrased. For example, the addition of the embedding $\mathbf{w}_{\text{royal}}$ aligns the distribution of $P(W|w_{\text{man}})$ to that of $P(W|w_{\text{king}})$ where W represents any and all the words in the vocabulary. The operation *contextualises* word *man*. The next section creates a paraphrase of expansionary and contractionary keywords (cf. equations (5) and (6)).

What does it then mean to subtract a vector embedding? Where addition contextualises a word or narrows its context, vector subtraction will *broaden* the context of a token. Bear in mind that subtraction is just an addition of a vector oriented in the opposite direction. In this sense, a subtraction will discontextualise a word, by removing the context that the subtracted word is described by. See, for example, Ethayarajh et al. (2019) who develop this intuition. Alternatively, one can also understand vector embedding subtraction as in Allen et al. (2019), by thinking about a difference in distributions. The authors show that a vector subtraction results in an embedding that is representative of the Kullback-Leibler divergence between the co-occurrence distributions of the two words. Difference between embeddings is, therefore, a measure of meaningful semantic change between two words. The next section subtracts the meaning of the contractionary paraphrase from the expansionary one to isolate the context that pertains to only one of these paraphrases (cf. equations (7)).

It can then be shown, as Ethayarajh et al. (2019) did, that any analogical relationship approximately satisfies, as

[27] cf. Table 1. There are almost 300'000 unique unigrams. This would be a space of dimensionality $\mathbb{R}^{300000 \times 300000}$.

[28] The informed reader will once again notice the similarity of the word vectorisation algorithms to the Principal Component Analysis (PCA).

Pennington et al. (2014) have conjectured in their original paper, the following approximate equality

$$\mathbf{w}_{word=a} - \mathbf{w}_{word=b} = \frac{P(W|a)}{P(W|b)} \approx \frac{P(W|x)}{P(W|y)} = \mathbf{w}_{word=x} - \mathbf{w}_{word=y} \tag{4}$$

This could be the canonical example $\mathbf{w}_{king} - \mathbf{w}_{queen} \approx \mathbf{w}_{man} - \mathbf{w}_{woman}$ or equivalently $\mathbf{w}_{king} - \mathbf{w}_{man} \approx \mathbf{w}_{queen} - \mathbf{w}_{woman}$ and $\mathbf{w}_{king} + \mathbf{w}_{woman} \approx \mathbf{w}_{queen} + \mathbf{w}_{man}$. If the difference/change in the 'degree' of royalty between man and king is the same between queen and woman, the analogical pair woman and queen is easy to be found because it bears the same semantic difference. The capability of embeddings to capture analogies of language will be used to make the narrative lexica derived in the next section more *sentimental* (cf. equation (8)). Figure 8 highlights the analogies in the space. For further discussion of the linear substructures, particularly the analogical relationships in the vector spaces, kindly refer to Appendix E.

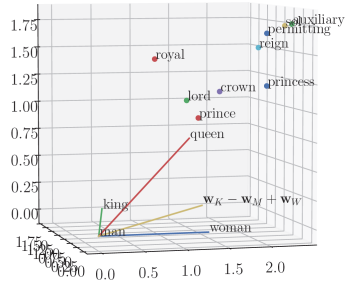


Figure 1. The relative locations of word embeddings for the analogy "*man* is to *king* as *woman* is to ..?". The closest embedding to the linear combination $\mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ is that of *queen*. We explain why this occurs and interpret the difference between them.

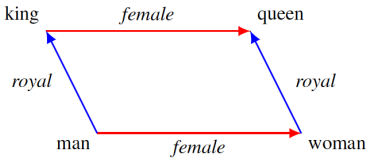


Figure 1: The parallelogram structure of the linear analogy *(king,queen)::(man,woman)*. A linear analogy transforms the first element in an ordered word pair by adding a displacement vector to it. Arrows indicate the directions of the semantic relations.

Figure 8: Analogical Relationships in a Vector Space: Taken from Allen and Hospedales (2019) (left) and Ethayarajh et al. (2019) (right). The chart is further discussed in Appendix E

4.4 Construction of the Indices

4.4.1 Lexicons: Expansionary and Contractionary Narratives – Vectors

To operationalise the ideas of Firth (1957), his distributional hypothesis, and utilise vector space models in the context of business cycles, I derived the following approach. The GloVe vector space model is estimated on the entire corpus of business cycle-related news reporting, exactly as outlined in points 1.–6. in Section 4.3.2. In the resulting space, a paraphrase for both the expansionary and contractionary cycle is created by summing the vectors of all expansionary and contractionary business cycle keywords respectively. Based on the literature reviewed in 4.3.2 and Appendix E, the summation gives us an embedding which should be located geometrically closest to contextual words which are common to all of these business cycle words, and thus bear the semantic meaning of all these business cycle keywords. The resulting embedding represents, in fact, a hypothetical word that bears an average of the meaning of these keywords. Specifically, observe the following sums of vector embeddings:

$$\mathcal{W}^{\text{contraction}} = \mathbf{w}_{\text{bust}} + \mathbf{w}_{\text{crisis}} + \mathbf{w}_{\text{contraction}} + \mathbf{w}_{\text{depression}} + \mathbf{w}_{\text{downturn}} + \mathbf{w}_{\text{recession}} \tag{5}$$

$$\mathcal{W}^{\text{expansion}} = \mathbf{w}_{\text{boom}} + \mathbf{w}_{\text{expansion}} + \mathbf{w}_{\text{recovery}} + \mathbf{w}_{\text{revival}} + \mathbf{w}_{\text{upturn}} + \mathbf{w}_{\text{prosperity}} \tag{6}$$

where $\mathcal{W}^{\text{contraction}}$ and $\mathcal{W}^{\text{expansion}}$ are paraphrases of the combinations of the contractionary and expansionary keywords. The individual vectors are the embeddings derived by the GloVe model – for example, $\mathbf{w}_{\text{recession}}$ is the embedded word vector of the word *recession*. The use of calligraphic w should emphasise that these are vectors resulting from a sum of other vectors. We could stop here and create a list of words that are deemed to be

closest (by cosine similarity) to this hypothetical word given by the two paraphrases, $\mathcal{W}^{\text{contraction}}$ and $\mathcal{W}^{\text{expansion}}$. However, one should note that it could still be that many of the words that are commonly found around the hypothetical paraphrase of contractionary and expansionary keywords are similar. For example, it is reasonable to expect that the word *economy* would be found close to both of these vectors. This semantic similarity between the two paraphrases means that just by the operation in (5) and (6), we cannot truly differentiate between the vocabulary that is often used near the words in the respective sums. $\mathcal{W}^{\text{contraction}}$ and $\mathcal{W}^{\text{expansion}}$ are, after all, *similar in some sense*. See Figure 23 for the closest words to both these paraphrases. Based on the discussions in Section 4.3.2 and Appendix E, it would be a reasonable next step to build a difference between these two vectors:

$$\mathcal{W}^{\text{expansion}} - \mathcal{W}^{\text{contraction}} \quad \text{or} \quad \mathcal{W}^{\text{contraction}} - \mathcal{W}^{\text{expansion}} \tag{7}$$

Given the argumentation outlined in Section 4.3.2 and Appendix E, the first difference will lead to an embedding that represents a hypothetical word whose probability distribution over contextual words (encoded in the coordinates) removed the context that was commonly found next to the vector for contraction, $\mathcal{W}^{contraction}$. Literally, the operation subtracts the semantic meaning encoded in $\mathcal{W}^{contraction}$ from that encoded in $\mathcal{W}^{expansion}$. Intuitively, the resulting coordinates place this new vector close to words that were found relatively often near the expansionary words, *but were not* found relatively often near the contractionary words. Again, we could stop here and look for the closest words to each of these two resulting vectors which would give us a list of words that were found to be related to only either $\mathcal{W}^{expansion}$ or $\mathcal{W}^{contraction}$[29]. However, it could be argued, as Shiller (2017) strongly emphasises, that a story motivates and connects behaviour to *"deeply felt values and needs"*. Based on Section 2.2, for a story to truly classify as a narrative, people need to attach certain values, judgments, or basically a degree of subjective emotion to it (cf. the fourth proposition of Shiller (2019) there). Therefore, to uncover words that have an emotional significance to them – in terms of business cycles here – we would ideally want to move the resulting vectors from equation (7) closer to such words with emotional significance[30]. For this purpose, the review of word vector analogies from above is utilised. First, find a word that bears a clear sentimental value. In sentiment analysis, the words *positive* and *negative* are often used to describe sentiments of words. There is no a priori reason for choosing exactly *negative* and *positive* as two sentimental words, beyond that they are on themselves *unequivocal* in their sentimental meaning. *good* and *bad* would be fitting candidates as well, and using them will result in a very similar vocabulary of closely associated words. Since these words are also often found in the news articles, they should be appropriate for use as the model has seen them in many contexts. Suppose we believe that the word *positive* corresponds to the sentiment representative of the expansionary stage of the cycle, and *negative* to the sentiment of the contractionary stage of the cycle. In that case, based on the arguments from the previous section, one should expect the following approximate vector difference equality:

$$\mathcal{W}^{\text{expansion}} - \mathcal{W}^{\text{contraction}} \approx \mathbf{w}_{\text{positive}} - \mathbf{w}_{\text{negative}} \tag{8}$$

In other words, if one is willing to accept the assumption that the semantic difference along a 'cycle' dimension is approximately similar to the semantic difference along a 'sentimental' dimension[31], then it is reasonable to believe that the following operations will lead to a vector that is close to words being semantically related to expansion and contraction individually, and bear a degree of sentimental value:

$$\mathcal{W}^{\text{expansion}} - \mathcal{W}^{\text{contraction}} + \mathbf{w}_{\text{negative}} = \mathcal{W}^{\text{ES}} \tag{9}$$

$$\mathcal{W}^{\text{contraction}} - \mathcal{W}^{\text{expansion}} + \mathbf{w}_{\text{positive}} = \mathcal{W}^{\text{CS}} \tag{10}$$

[29] See Figure 24 in Appendix G for the two lexicons which would emerge from the closest words to the vector defined by equation (7). It can be seen that words such as *economy*, that often co-occurred with words in both $\mathcal{W}^{\text{contraction}}$ and $\mathcal{W}^{\text{expansion}}$ disappeared.

[30] Another problematic issue emerges from equation (7). The vectors are too close to the original words in $\mathcal{W}^{\text{contraction}}$ and $\mathcal{W}^{\text{expansion}}$. Many of the closest tokens are thus merely bigram combinations of these keywords. This can also be seen in Figure 24.

[31] This is a testable assumption. In fact, the Euclidian distance between $\mathcal{W}^{\text{expansion}}$ and $\mathcal{W}^{\text{contraction}}$ is 0.85. Between $\mathbf{w}_{\text{positive}}$ and $\mathbf{w}_{\text{negative}}$ it is 0.68. The Euclidian distances between $\mathcal{W}^{\text{expansion}}$ and $\mathbf{w}_{\text{positive}}$ and vice versa are almost identical. The parallelograms as in Figure 9 do approximately exist. Using *good* and *bad* results in an even exacter parallelogram.

$$\{\mathbf{w}^{j,\mathrm{ES}} \approx \mathcal{W}^{\mathrm{ES}}\} : \text{Expansionary Sentiment Lexicon}$$
$$\{\mathbf{w}^{k,\mathrm{CS}} \approx \mathcal{W}^{\mathrm{CS}}\} : \text{Contractionary Sentiment Lexicon}$$

where $\mathcal{W}^{\mathrm{ES}}$ and $\mathcal{W}^{\mathrm{CS}}$ denote the expansionary and contractionary narrative vectors. Their location in the vector space is used to derive the two lexicons: $\{\mathbf{w}^{j,\mathrm{ES}}\}$ and $\{\mathbf{w}^{k,\mathrm{CS}}\}$. Notice that these are sets of vectors so that $\{\mathbf{w}^{j,\mathrm{ES}}\} = \{\mathbf{w}^{1,\mathrm{ES}}, \mathbf{w}^{2,\mathrm{ES}}, ..., \mathbf{w}^{J,\mathrm{ES}}\}$ and $\{\mathbf{w}^{k,\mathrm{CS}}\} = \{\mathbf{w}^{1,\mathrm{CS}}, \mathbf{w}^{2,\mathrm{CS}}, ..., \mathbf{w}^{K,\mathrm{CS}}\}$ with each embedding j, k representing a word from the vocabulary $\mathcal{V}$. The two sets of vectors that were found close – similar – to vector $\mathcal{W}^{\mathrm{ES}}$ and $\mathcal{W}^{\mathrm{CS}}$ respectively represent tokens from the vocabulary that were found most associated with the expansionary and the contractionary stages of the cycle, and hence embody the respective narratives. The collection of these words is the basis for the contractionary and expansionary lexicons used to construct the Relative Sentiment Index.[32].

More specifically, to isolate the most similar words after deriving $\mathcal{W}^{\mathrm{ES}}$ and $\mathcal{W}^{\mathrm{CS}}$ in (9) and (10), cosine similarity between them and the other vectors in the space is calculated (cf. also the generic definition of cosine similarity in Appendix E, equation (17)). The following statistic is computed for all embeddings $\mathbf{w}$ in the vector space:

$$S(\mathcal{W}^{\mathrm{CS}}, \mathbf{w}) = \frac{\sum_{i=1}^{n} \mathcal{W}_i^{\mathrm{CS}} w_i}{\sqrt{\sum_{i=1}^{n} (\mathcal{W}_i^{\mathrm{CS}})^2} \sqrt{\sum_{i=1}^{n} (w_i)^2}} \tag{11}$$

where the i in $\mathcal{W}_i^{\mathrm{CS}}$ and w_i stands for an element of the word vectors $\mathcal{W}_i^{\mathrm{CS}}$ and w_i that represent the words – or rather their meaning. Note that $\mathcal{W}^{\mathrm{CS}}$ does not represent an actual word, only a hypothetical, but the information encoded in its coordinates bears information on what I term contractionary narrative sentiment. Subsequently, a cut-off in the metric is chosen to isolate the most related terms. I shall use 0.3 as a cut-off to extract all word vectors $\mathbf{w} \in \mathcal{V}$ that are sufficiently close to either the expansionary or contractionary narrative word embedding. Equivalently, for all vectors j, k in sets $\{\mathbf{w}^{j,ES}\}$ and $\{\mathbf{w}^{k,CS}\}$, the following conditions hold: $S(\mathbf{w}^{j,\mathrm{ES}}, \mathcal{W}^{\mathrm{ES}}) > 0.3$ and $S(\mathbf{w}^{k,\mathrm{CS}}, \mathcal{W}^{\mathrm{CS}}) > 0.3$. Therefore, after the operation in equation (11), only vectors $\mathbf{w}$ where

$$S(\mathcal{W}^{\mathrm{CS}}, \mathbf{w}) > 0.3 \tag{12}$$

are kept for the contractionary lexicon, $\{\mathbf{w}^{k,\mathrm{CS}}\}$. The analogous calculation is done for $\mathcal{W}^{\mathrm{ES}}$ resulting in $\{\mathbf{w}^{j,\mathrm{ES}}\}$. Together, the words represented by vectors $\{\mathbf{w}^{j,ES}\}$ and $\{\mathbf{w}^{k,CS}\}$ make up the expansionary and contractionary narrative lexicons. The resulting dictionaries describing each of the two business cycle sentiments can be found in Tables 13 and 14 in Appendix G.

The model is estimated with *all* articles pertaining to business cycles in the corpus. To provide further robustness to the results, and stronger evidence of the predictive power of the resulting index, the vector space model will also be re-estimated with an early subset of articles, so that future news talk cannot influence the creation of lexicons. However, the reader should note that since the thesis uses both U.S. and European news over thirty years, the estimated lexicon should not be susceptible to regional or timely linguistic outliers. As the reader is welcome to observe in Figure 10 and Tables 13 and 14, the great majority of words identified are common words that could have been used in any time and in different contexts, rather than only when talking about a particular macroeconomic event.

The cosine similarity cut-off could be adjusted. There is an underlying signal-to-noise trade-off in choosing this cut-off. If higher, smaller lexicons emerge, creating possibly a more exact narrative proxy, but fewer words to count, and therefore a potentially weaker signal. If lower, more numerous lexicons emerge, where the words might

[32]To get a better grasp of the vector operation in (9) and (10), one can think of the canonical analogy from section 4.3.2: a man is to woman, as a king is to what? In the same way we can ask, contraction is to negativity as expansion is to what exactly? And vice versa for expansion. Admittedly, we could have for example asked, contraction is to anxiety as expansion is to ...? – or vice versa. We are, however, limited by the corpus, where words like anxiety or enthusiasm are used only sporadically.

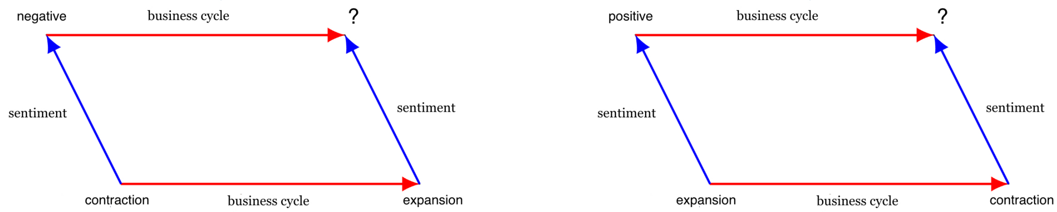


Figure 9: Illustratory Parallelogram Spanned in the Vector Space. Left: Expansionary Lexicon. Right: Contractionary Lexicon. The respective lexicons emerge as two collections of words that the vector space considers most fitting to take the place of the question marks in the analogy above. Source: Own Diagrams

be only weakly related to business cycles, and hence a noisier index emerges. Depending on the pre-specification, this dictionary by no means has to be limited to unigrams. In this study, unigrams and bigrams shall be used for the narrative lexicons. Therefore, *token* instead of *word* or *term* was used previously, to emphasise that the vectors do *not* represent merely unigrams. In the case of this thesis, the co-occurrence matrix consisted of a symmetric context window of ten tokens to the left and ten tokens to the right (20 context words in total were evaluated). See the hyperparameters set in Table 2.

Figure 10: Final Narrative Sentiment Lexicons: Expansionary (left) and Contractionary (right). Size by Cosine Distances to $\mathcal{W}^{\text{ES}}, \mathcal{W}^{\text{CS}}$.

Note how the resulting lexicons could not only be interpreted as bearing a certain business cycle narrative sentiment but be an expression of animal spirits in general. Nyman et al. (2018) have laboriously crafted a lexicon to capture their conception of animal spirits which they see expressed in words resembling the meaning of excitement and anxiety, or equivalently, of approaching versus of avoiding risk. Several of the words that they manually collected naturally result from this automatised method[33].

4.4.2 Index 1: Expansionary, Contractionary and Relative Business Cycle Sentiment

The Relative Sentiment Index is a result of the difference between the Expansionary Sentiment time series and the Contractionary Sentiment time series. These two series are constructed as proportions of counts of tokens stemming from the expansionary and contractionary lexicons to the count of all tokens used in the economic

[33] Nyman et al. (2018) did not publish their entire lexicon. They tabulated 40 random 'excitement' and 40 random 'anxiety' words. Out of these 80, 10 can be found in either the same or similar grammatical form in the narrative lexicons identified here.

Table 2: Hyperparameters: GloVe Model Estimation

'Degrees of Freedom': Chosen Hyperparameters		
Type of Choice	Choice	Reasoning
Degree of N-grams considered	2	Given the size of the business cycle corpus (around 60'000 articles), it might have been counter-productive to examine N-grams of a higher order. Introducing higher-order N-grams would mean more semantic relationships to learn, given the same amount of information. I have not found learning on trigrams to improve the index. Learning on bigrams as compared on unigrams, however, did make the Relative Sentiment Index more interesting.
Dimensionality of token embeddings	100	Experimenting with changes of this hyperparameter did not appear to result in better performance – especially not if higher dimensionality was chosen. This finding is roughly in line with the conclusions of Pennington et al. (2014), who also argued that there are diminishing marginal returns to accuracy when increasing the dimensionality of embeddings. This value is in line with the pre-trained embeddings available online, and in the range used in the literature. Based on the review of the algorithm, it could be advantageous not to choose too many dimensions. Not only because it costs additional time to estimate the model, but also because the dimensional compression is where interesting patterns are derived.
Maximum co-occurrences considered	100	The choice corresponds with choices in the original paper of Pennington et al. (2014). Therefore, it is rather high in the context of the smaller corpus in this study.
Number of iterations in model estimation	200	In the original paper of Pennington et al. (2014), it was shown that the algorithm converges quickly to the optimal solution – this value should be enough.
Context window considered	10 (symmetric)	Possibly the most important hyperparameter. Ten words to the left and ten words to the right of each target word when learning its embedding were considered. For larger windows, semantic rather then syntactic relationships are discovered. Twenty words are many; the focus is thus on semantics. The algorithm under-weights words that are relatively more distanced from the target.
Vocabulary cut-off for learning	Occurrence in at least 50 news articles	Only token embeddings of terms occurring at least in 50 different articles are considered. This is to ensure less noise in learning. Experience has shown that larger values lead to better and less noisy results. With 60000 articles, this does not seem too restrictive.

expectation news in each period[34]. The time series are created in monthly intervals – token proportions for each month of the news reporting are calculated. The following three series are constructed based on Corpus 2 (Economic Expectations): *ES* (Expansionary Sentiment), *CS* (Contractionary Sentiment) and *RS* (Relative Sentiment). All underlying news articles are related to the United States. The series are defined as:

$$ES_t = \frac{\text{Count}(\text{Token} \in \{\mathbf{w}^{j,ES}\})_t}{\text{Count}(\text{All tokens})_t} \tag{13}$$

$$CS_t = \frac{\text{Count}(\text{Token} \in \{\mathbf{w}^{k,CS}\})_t}{\text{Count}(\text{All tokens})_t} \tag{14}$$

$$RS_t = ES_t - CS_t \tag{15}$$

where *token* instead of words was used to emphasise that these are also bigrams, or two consecutive words. See the complete list of terms in Tables 13 and 14. The RS_t series will henceforth only be referred to as the Relative Sentiment (uppercase) or the Relative Sentiment Index.

[34] An alternative to simple counts of the lexical words would be to make use of some relative measure of semantic association. A natural candidate for this would be a sum of cosine distances of the cut-off words in the articles. This approach was attempted, however, led to only marginal improvements.

4.4.3 Index 2: Narrative Consensus and Shannon's Entropy

The second aim is to create a measure of narrative consensus in the news articles, understood as news being focused on a specific aspect – topic – versus being very broad in the content, discussing a wide variety of issues. This focus will be proxied with the topic distributions derived from LDA over time. After constructing the LDA model, a measure of uncertainty – Shannon's entropy – is applied to create this indicator. The analysis here included will be *only* performed on Corpus 2 (economic expectations). The construction of this index is independent of the RS_t time series created above.

The optimal number of topics is determined with the help of R package *ldatuning* that implements several algorithms to find this optimal number of latent structures. The package implements the work of Deveaud, Sanjuan, and Bellot (2014), Cao, Xia, Li, Zhang, and Tang (2009) and Arun, Suresh, Veni Madhavan, and Narasimha Murthy (2010). See Figure 30 in Appendix K for a graphical representation. Subsequently, the model with the optimal amount of topics based on these criteria is chosen. For each article in the corpus, the LDA model generates two metrics – β and γ as outlined in Section 4.3.1 and Figure 5. Whereas β gives a probability distribution over words of a topic, γ gives a probability distribution over topics for a document. See Figure 31 in Appendix K for a graphical representation of the β-distribution of several topics. Since each period in time consists of collections of documents, each having a distribution over topics, each period in time will also have a distribution over this latent structure defined by γ. The hypothesis here is that if this distribution is rather concentrated on a relatively small number of these topics, the discourse is concentrated too, its diversity is low, and so economic beliefs are also relatively concentrated and homogeneous[35]. Note that the notion of what diversity and concentration means is inseparable from how the LDA defines the topics. The character of the topics might as well be unintuitive to how the researcher might be thinking of diversity of discourse.

Nevertheless, we want to have a measure of this diversity or concentration. The natural candidate for this is Shannon's entropy. When a given distribution is relatively narrow, concentrated on a small subset of outcomes, entropy will be small as the uncertainty of the outcome is small as well. If a distribution is rather wide, dispersed with numerous peaks or in the limit, uniformly distributed, the entropy, as well as uncertainty, is maximised. The γ-distribution derived with LDA, when aggregated at monthly intervals, is used to calculate Shannon's entropy.

The entire structure of the estimated γ-distribution is incorporated into the Entropy Index as follows. For each period (one month), the following sum, aggregating the gammas of all articles for each topic i in a time period t, $\sum_{d=1}^{D} \gamma_{d,i,t}$, is calculated. This number is then divided by the number of articles D that period. Doing this for every i, a discrete probability distribution over topics in that period is generated, denoted γ_t. The probability of topic i, in period t is then $\gamma_{i,t}$. These probabilities are used to construct the Narrative Consensus Index as:

$$\text{Entropy}_t = -\sum_{i=1}^{k} \gamma_{i,t}\, log(\gamma_{i,t}) \tag{16}$$

The mechanism works as follows: Entropy will be maximised if articles are found to be uniformly distributed over all topics. Any concentration (consensus) will result in lower entropy. Therefore, the lower the entropy, the larger the presumed consensus and vice versa. Nyman et al. (2018)[36] have shown that Narrative Consensus increased

[35] A case to the contrary might be a major event (e.g., an emerging pandemic) which renders the discourse to be concentrated on a single (or small amount of) topic(s). It is not ex-ante given that a model such as LDA, especially if estimated on a thirty years period as here, will recognise Covid or 'pandemic' as a specific individual category/topic. Therefore, there is no a priori reason to believe that the emergence of a pandemic will result in less diverse economic discourse. The exact opposite could be the case since the pandemic touches so many aspects of the economy, that momentarily, the entire universe of economic topics will be discussed.

[36] As an excursion, an alternative would be to partly use the approach of Nyman et al. (2018) and denote each article as carrying only one topic. Remember that their method assigned each document only a single, most representative topic. In such a case, Shannon's entropy must be calculated as

$$\text{Entropy}_t^{\text{BOE}} = -\sum_{i=1}^{k} \frac{n_{i,t}}{N_t}\, log\left(\frac{n_{i,t}}{N_t}\right) = log(N_t) - \frac{1}{N_t}\sum_{i=1}^{k} n_{i,t}\, log(n_{i,t})$$

where n_i stands for the number of articles (during a time period) of topic i and N is the number of all articles that month. It

before the recent crisis. Similar to their finding, Larsen and Thorsrud (2019a) identified increasing sparsity of topics contributing to their index during expansions, which could be interpreted as high consensus (low entropy) expansions and increasing entropy (lowering consensus) pre- and after crises. In comparison to this thesis, the former did not use LDA to create the topics, and the latter used a dynamic variable in their latent threshold model of the macroeconomy to approximate the narrative consensus.

4.5 Evaluation: Macroeconomic Time Series and Statistical Testing

One crucial methodological question remains: How can we evaluate the usefulness of the Relative Sentiment Index and the Entropy Index for capturing and possibly predicting business cycles? Two components needed for this goal are introduced below – comparative time series and statistical tools.

Firstly, we will need to compare the Relative Sentiment and Entropy Indices to other macroeconomically relevant time series. An obvious comparison candidate is a measure of the Gross Domestic Product. GDP is what most people associate with business cycle measurement. The analysis here will make use of U.S. quarter on quarter, seasonally adjusted, annualised GDP growth (nominal and real)[37]. The connection between GDP (QoQ) growth and the business cycle is relatively straightforward. For example, if this measure of GDP growth is positive and high, nobody would dispute we find ourselves in an expansionary stage of the cycle. Conversely, if the GDP growth decreases for a couple of consecutive quarters, or at latest when it turns negative, contraction ensues[38]. The text-based indices should thus be able to capture similar patterns and trends as the quarterly, seasonally adjusted, GDP growth. The source for the GDP time series is, as is the case for every comparative time series used, the database of St. Louis FED called FRED. I extract both nominal and real QoQ GDP growth for the U.S. starting Q1 1990 and spanning until Q2 2020 (U.S. Bureau of Economic Analysis, n.d.-a, n.d.-b).

Apart from the GDP, additional series deemed interesting on the grounds of their macroeconomic relevance, connection with the sentiment, beliefs about the future of the macroeconomy, and the perception of the general level of macroeconomic uncertainty, are collected. These are the Economic Policy Uncertainty Index created by Baker et al. (2016), *EPU*, the CBOE Volatility Index or *VIX*, commonly referred to as the 'fear index', and *UMCSENT*, the University of Michigan Consumer Sentiment Index. These are all available at monthly time intervals, from January 1990 until Apr 2020, from the FRED platform (Baker, Bloom, & Davis, n.d.; Chicago Board Options Exchange, n.d.; University of Michigan, n.d.). The time series used are described in the Table 3.

Table 3: Comparative Time Series Database: Descriptive Table

Time Series Descriptives							
Constructed Text-Based Time Series	Begin	End	Tot Periods	Min	Mean	Max	SD
Expansionary Sentiment	09/1987[39]	04/2020	392(364)	0.0015	0.0081	0.0152	0.0023
Contractionary Sentiment	09/1987[39]	04/2020	392(364)	0.0010	0.0081	0.0289	0.0035
Relative Sentiment	09/1987[39]	04/2020	392(364)	-2.498e-02	-3.663e-05	8.526e-03	0.0042
Entropy	09/1987[39]	04/2020	392(364)	2.312	3.596	4.047	0.2883
Comparative Time Series	Begin	End	Tot Periods	Min	Mean	Max	SD
GDP, nominal, QoQ Growth (U.S.), qtr	Q1 1990	Q2 2020	122	-33.3	4.2	10.2	4.3
GDP, real, QoQ Growth (U.S.), qtr	Q1 1990	Q2 2020	122	-31.7	2.2	7.5	3.9
VIX (CBOE Volatility Index), mth avg	01/1990	04/2020	364	10.13	19.31	62.64	7.76
UoM Consumer Sentiment (UMCSENT), mth	01/1990	04/2020	364	55.30	87.59	112.00	12.32
Economic Policy Uncertainty (EPU US), mth	01/1990	04/2020	364	57.20	109.49	283.15	35.51

was found that using LDA, and extracting the single representative topic of each article resulted in noisy time series that was not informative and less interesting in the evaluations below.

[37] Alternative approach would be to decompose the level time series of GDP into a cyclical and a trend component as commonly done in the literature. The alternative was rejected on the grounds of not wanting to introduce further degrees of freedom to the analysis. With GDP decomposition, a researcher could greatly influence what one defines as the cyclical component.

[38] The concept of technical recession is another clear example. An economy is deemed to be in recession if GDP growth has been negative for two consecutive quarters.

[39] Only observations starting 01/1990 will be used for the econometric analysis below. This restriction was deemed appropriate,

Regarding the statistical tools, we need methods that can evaluate the indices in three crucial ways. In their relationship to the current macroeconomic activity, their predictive capacity, and their ability to exhibit breaks pre-business cycle turning points. To the former two purposes, cross-correlation functions, simple regressions (incl. in- and out-of-sample predictions) and Granger causality are examined. For the latter, structural break analysis is employed.

In terms of the correlation functions, we want to establish correlations of the text-based indices with leads of GDP growth and the comparative series. For a brief overview of correlation functions, please refer to Appendix H.2. Regarding linear regressions, we are interested in evaluating predictive success – particularly in terms of the sign, trend and level of GDP growth forecasts. Significance of the coefficients in various models, R squared values, and F-statistic, are also important. Any specification where GDP is regressed merely on the text-based index suffers from potentially substantial omitted variable bias. This issue should, however, not discourage us, since proving causality is not the aim. The regression and forecasting procedures are discussed in Appendix H.3. Subsequently, Granger causality is estimated via the approach of Toda and Yamamoto (1995) to establish whether the text-based indices are useful to forecast GDP growth beyond the information contained in the GDP series. The concept, as well as the testing procedure, are outlined in Appendix H.4. Lastly, multiple structural breaks are identified with the approach of Bai and Perron (2003) to evaluate the hypothesis that the text-based indices can predictively capture business cycle turning points and that their major shifts coincide with important macroeconomic developments. The procedure identifies the shifts via the most pronounced changes to the mean of the series in a time segment. Appendix H.5 offers an insight into the technique. In the following paragraphs, scaling and filtering is used to evaluate the patterns of the indices visually. The procedures used are outlined in Appendix H.1.

5 Results: Measuring Business Cycle Sentiment and Consensus

After carefully explaining the methods and the underlying corpora, we can proceed to examine results. The section commences with a descriptive discussion of the indices created – the Relative Sentiment and Entropy – and proceeds to evaluate their success in a variety of predictive tasks. Descriptive discussion of the visual patterns of the text-based time series is provided in the following two sections. The trend patterns which these series exhibit were found interesting, and could provide clues into how beliefs and narratives evolve during the business cycle. Some of the results have been deferred to Appendices J–N.

5.1 The (Narrative) Relative Sentiment Index

Consider first the questions of what the narrative business cycle sentiment index actually captures. What information do the lines in Figures 11 and 12 convey exactly? As can be seen in the equations (13) through (15), the vertical axis – the value of the indices – is plotting proportion of specific terms, unigrams and bigrams, out of all terms in economic news over time. These terms (full list in Tables 13 and 14) have been found to be

1. Related to key business cycle words through their similar contexts (equations (5) and (6))
2. Related relatively more strongly to *only* either expansionary or contractionary business cycle keywords (equations (7))
3. Have a certain sentimental value for the economic agents when it comes to the respective business cycle stage (equations (9) and (10))

Notice that if any of these three were not true, surely, it would be more difficult to argue that the information

especially because Relative Sentiment and Entropy Indices are substantially noisy pre-1990 as a result of only a few observations – news articles. Another reason is inconsistency in comparative series. VIX has followed different methodology prior 1990. Structural break analysis is an exception – entire corpus is used, mainly to capture the behaviour pre-1990–1991 recession.

captured by the index is of a narrative nature[40]. Remember that in Section 2.2, we saw how narratives are defined as "a story or representation used to give an explanatory or justificatory account of a society, period, etc.". It can, therefore, be argued that what separates a story from a narrative, is the attachment of a specific meaning. By construction, if we accept the proposition of Firth (1957), then our vector space model captures the meaning of words, and therefore the incidence of the words in the lexicon that are tracked in the index indicates how 'contractionary' and how 'expansionary' the news reporting is. As such, the index will measure how *semantically close* news reporting is to the meaning of contraction and the meaning of expansion at each point in time.

Let us take a closer look at its evolution over time. Observe first the four lines plotted in Figure 13. The black line plots the CS_t (contractionary) and grey line the ES_t (expansionary) time series. The coloured lines, blue and red, are HP-filtered series of these two respectively. It is important to stress that the absolute level of both of the indices bears little to no meaning – if the lexica were made larger, such as by decreasing the cosine similarity cut-off in equation (12), the values would increase by construction. It is merely the relative level, the trend, and the changes over time which are interesting and could be interpreted. The relative level is plotted in Figure 12 and constitutes what I refer to as the Relative Sentiment Index. Observe also the grey rectangles in both figures below. These refer to the official U.S. recession periods, as defined by National Bureau of Economic Research (n.d.). It can be clearly seen in Figure 11 that over these periods, the contractionary series has *always* reached its local maximum (both for the actual and for the HP-filtered series). The contrary applies to the expansionary series which has *always* reached its local minimum during these periods. Same goes for the Relative Sentiment Index in Figure 12. All crises have led to pronounced local minima in the RS_t Index, with the latest financial crisis being 'unbeaten' until the current Covid crisis. It can be seen that Covid led to relatively even more 'contractionary' writing in the economic expectation news than the financial crisis.

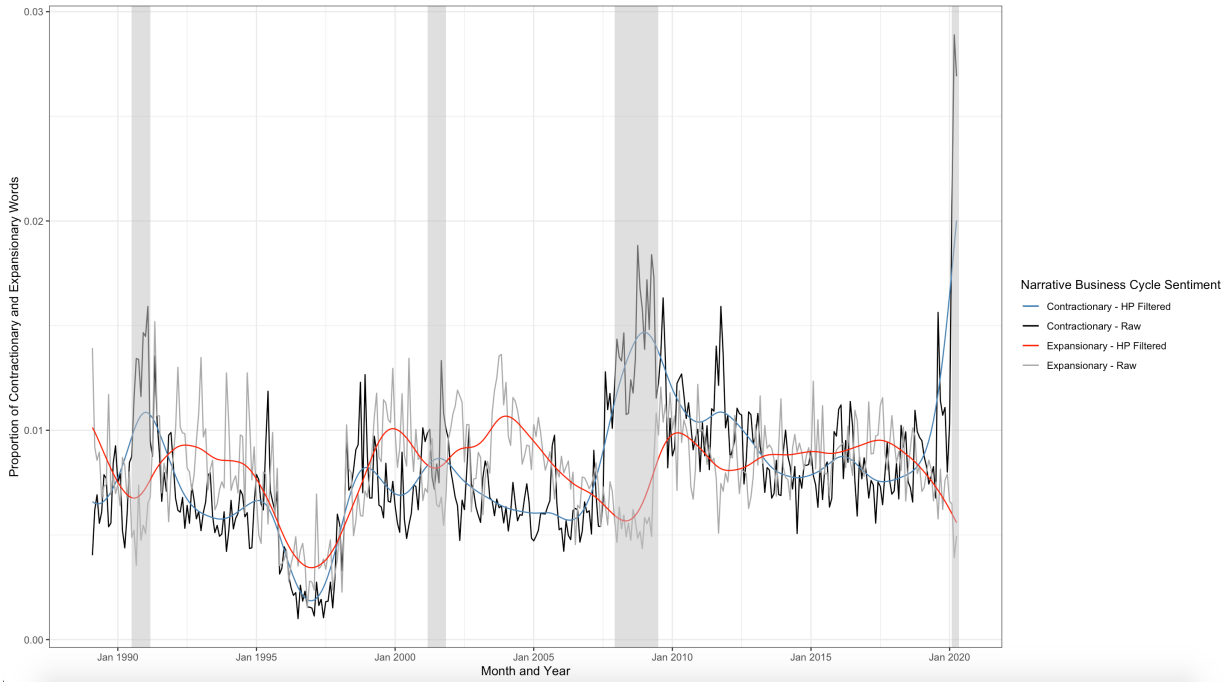


Figure 11: Expansionary and Contractionary Narrative Business Cycle Sentiment: Plotted Starting January 1989

What is perhaps more interesting are the long-term trends visible in the indices as well as the timing of these trends. To begin, examine Figure 11. There is a clear increasing (decreasing) trend in contractionary (expansionary) sentiment *before* each crisis. Perhaps the index can capture a cyclical transition in convictions about the state of

[40]Notice how it is important to clearly define what a narrative is to be convincing when it comes to arguing that an index such as this captures narrative aspects of economic thinking. This is, however, problematic on purely scientific terms. There is no unequivocal definition of narratives, and the existing definitions are even less clear when one wants to refer to economic narratives. Cf. Sections 2.2, 4.4.1 and 7 for arguments in favour of this index capturing economic narratives.

the economy. In the periods where the grey line has tended to decrease, and the black tended to increase, it could be claimed that agents noted that the economy is not in as solid a shape as they thought, and gradually updated their expectations towards pessimistic future outcomes. Notice too that these counter-moving trends of ES_t and CS_t are pronounced *only* before crises. There is considerable noise in the constructed indices, which could be explained with the relatively small size of the corpus on economic expectations news. As can be seen in Figure 1, there are often 100 or fewer news articles each month that the index is based on. Therefore, the variation can become substantial if there are only several 'outlier articles' containing many of the words from the two lexica. Because of this noisy aspect, it may appear challenging to draw conclusions about when a significant change in trend has occurred, or estimate when a convincing prediction could be made. Therefore, both a Hodrick-Prescott filter and a double exponentially smoothed series were added in Figure 12. The latter does not use future values in smoothing. A pronounced decreasing tendency can be seen before each crisis, albeit the picture is admittedly distorted by the fact that there are pronounced decreases at other times too. Notably, the second half of the 2000s (2006 – 2009) exhibits a pronounced pre-crisis decrease as well as the early 2000s and somewhat a less visible one around 1997. The former ensues shortly before the recent financial crisis. The latter two, it could be fairly well reasoned, correlate with the uncertainty around the dot-com bubble. Perhaps, many market participants expected a substantial economic contraction much earlier than it actually arrived[41].

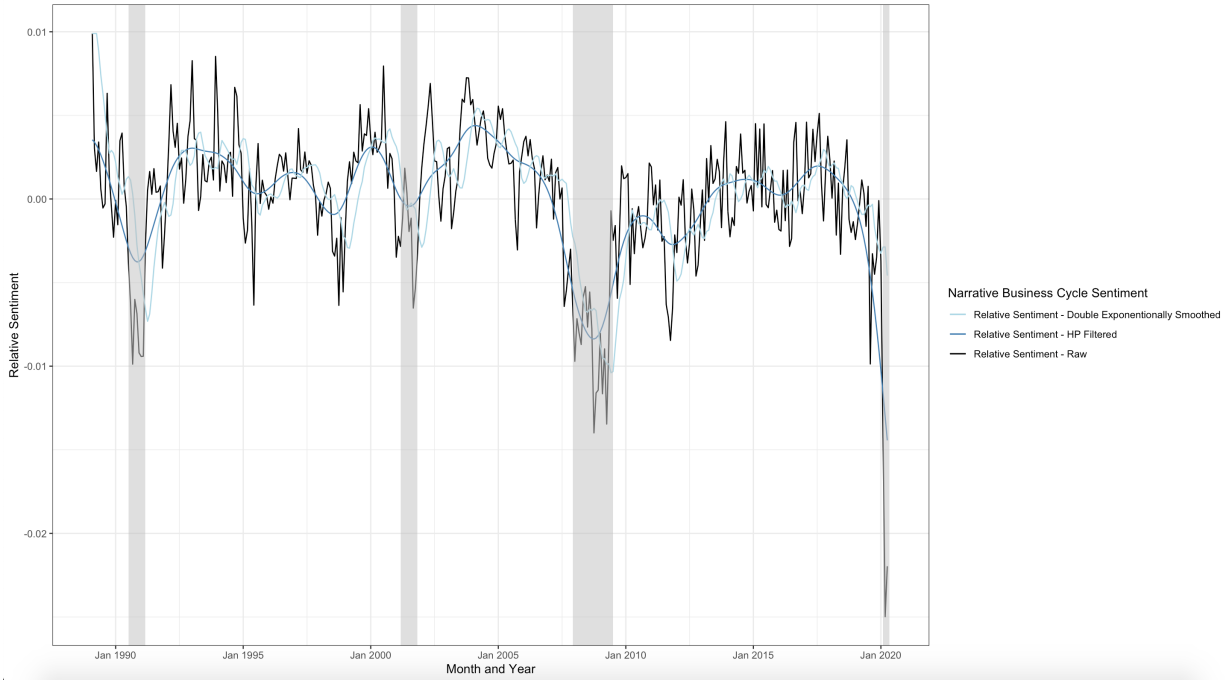


Figure 12: (Narrative) Relative Sentiment Index: Plotted Starting January 1989

Furthermore, notice that the revival of expansionary sentiment does not tend to come until just before the crises 'officially' ended, mostly even just afterwards. The expansionary lexicon could seemingly be improved – or perhaps it is more difficult to capture positive business cycle sentiment from news articles than it is to capture the negative. It could also generically be that both the narrative lexica constructed (Tables 13 and 14) include too many business cycle keywords. For example, the word *recovery* can be found in the expansionary lexicon but is likely to be used post-factum, when the economy is found to be truly recovered. Similarly, there is not much reason to use the word *crisis* until there is an actual economic crisis transpiring. Removing such 'post-factual' words from the index would be an interesting avenue for an extension, but has not pursued further in the thesis.

[41] As is commonly acknowledged, the dot-com bubble started bursting at the end of 1999 and continued into the year 2000. It led to the recession denoted by the grey rectangle in the early 2000s. The 1997 Asian financial crisis and the associated October 27, 1997 mini-crash might have perpetuated the economic expectations to flock around the contractionary sentiment. There is a local minimum around this period in the index. This pronounced minimum could be an example of a 'belief crisis' without an actual economic crisis immediately following. Ultimately, it was the next break in the series that transpired before an actual crisis.

Lastly, there should, arguably, exist some long term mean reversion in the Relative Sentiment Index. It seems unlikely and unnatural that sentiment, or even beliefs, as embodied by the words in both lexicons can remain being used heavily, or not at all, for longer periods of time. Perhaps the only potential example of such development that could be postulated is that a group of terms stops being used or/and is replaced by new terms because of linguistic shifts. Since the lexica have been trained on the entire time horizon of news articles, this should not be the case. It should, however, be stressed that the vector space model underlying the index would have to be regularly re-estimated to yield most accurate results. In Section 5.4, the reader can note that the general patterns and trends found here are remarkably stable, even with lexica derived from an early subset of news articles or when certain, perhaps visionary, words are removed from the vocabularies.

5.2 The (Narrative) Consensus: Entropy Index

Whereas the previous index dealt with measuring the prevailing business cycle sentiment in economic expectations, the second index deals with a different issue. The purpose of this exercise is to shed light on the degree of (dis-)agreement between economic agents in their assessment of the economy's future. The index measures the broadness of dialogue in the news on economic expectations in the U.S. As described in Chapter 4.3.1, a Latent Dirichlet Allocation model is estimated, featuring an optimal number of latent components, and Shannon's entropy of the topic distribution over document categories (the posterior γ-distribution) is calculated periodically.

The identifying assumption here is that discussing a wider variety of topics means that there is less agreement on the future state of the economy. Another assumption made implicitly is that the number of latent components did not change over the time horizon. These assumptions are not a panacea and will be discussed in the limitations part below. Note, however, that the stability of the topics does not mean that the wording representative of a specific topic cannot change over time. For example, a topic describable as 'technology news' might assign high probabilities to a broad spectrum of words that might come from different time periods such as *DVD* and *smartphone*. In this sense, it might even be of advantage to examine the broadness of focus with a stable categorical structure since the topics so created have a long-term relevance, and so comparisons over longer time horizons are perhaps more meaningful. It is then natural to think about a disproportionately large focus on a small subset of topics as a potential indication of the presence of an economic *bubble*[42], or at least suggesting that a specific subset of topics is of particular importance for expectation formation.

The Narrative Consensus Index is plotted in Figure 13. The index seems to exhibit less variation compared to its sentiment-capturing counterpart from the previous section, especially when it comes to examining trends. There appear to be two segments in the series where the index was either excessively volatile or exhibits a visibly lower mean. The trend is one of considerable decrease starting in 1995 that ultimately leads to a global minimum in mid-1997. This was the period when the dot-com bubble was forming, and the Asian financial crisis in combination with the October 1997 mini-crash precipitated in the U.S. Hence, it would make sense to argue that the index captured market tensions, perhaps even the emergence of an economic bubble in the run-up to the said mini-crash and dot-com bubble. Another mentionable period is the time before the latest financial crisis. The level of the entropy measure is again at its local minimum before the crisis ensued until it corrects again upwards as the crisis transpires. Also notable is the non-existence of any major decrease before the start of the Covid pandemic.

At this point, it makes sense to briefly remark on what lower and higher values of the narrative consensus measure do *not* mean. The reader would perhaps argue that it is reasonable to expect narrative consensus increase, and the Entropy Index, therefore, decrease, in the run-up to a crisis – and not increase as is the case here or in Nyman et al. (2018). Perhaps because the reader might think that aligning on the belief 'there will be a crisis' embodies a narrative consensus. Such a perspective is conceptually intuitive, albeit not in line with the workings of the model here. LDA does not uncover a topic that is representative of a crisis or a bubble; there is no 'economic crisis latent component' on which the agents would focus their discussion. There are only latent categories that could

[42]The behavioural economics literature that stipulated herding as a potential explanation for the emergence of economic bubbles would most likely agree with the interpretation of the $Entropy_t$ Index as a likely economic bubble index. Harmon et al. (2015), for example, focused on the co-movement of stocks as a measure of collective panic. Using a measure of narrative consensus could be a text-based counterpart to their endeavour.

be understood in more straightforward ways such as 'foreign policy', 'oil' or the 'stock market'[43]. Nevertheless, Larsen and Thorsrud (2019a) suggest that expansions are driven by a larger variety in economic discourse, whereas contractions are focused.

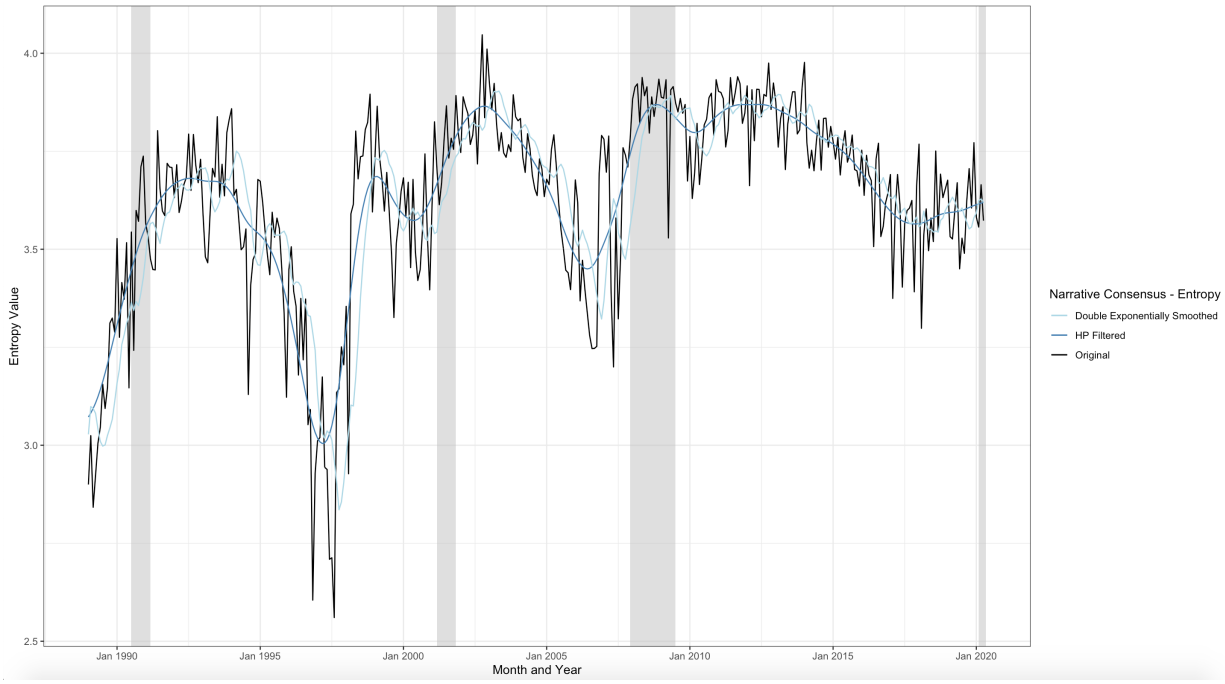


Figure 13: Narrative Consensus: Plotted Starting January 1989

5.3 Evaluating the Indices

Recall the title: Can We Predict Business Cycles With Natural Language Processing? This section and the evaluation analysis is broken into several parts that will provide different perspectives and insights helpful in answering this overarching question. First, the co-movement of the text-based indices and quarterly GDP growth in the U.S. will be examined. Is there a stable relationship between them? (Cross-)correlation and later cointegration of the series are examined. To gauge the predictive power of the indices, regression analysis, simple in-the-sample and out-of-sample predictions are performed. Finally, to see whether the indices can capture business cycle turning points, structural break tests are constructed.

For comparison and further robustness, many of the above tests are also evaluated in the context of the relationship of the constructed text-based indices to other similar indices such as the CBOE Volatility Index (VIX), University of Michigan Consumer Sentiment (UMCSENT) and Economic Policy Uncertainty Index (EPU). Section 4.5 elaborated on the relevance of these comparisons. For plots of scaled time series comparisons that are left out in the evaluation below, cf. Appendix J and Figures 25 through 29.

5.3.1 Correlation and Cross-Correlation Functions

In Figure 14, cross-correlation functions between RS_t and U.S. quarterly, seasonally adjusted, nominal GDP growth, *VIX*, *UMCSENT* and *EPU* are plotted. See Figure 32 in Appendix L for the corresponding metric for $Entropy_t$. The lags and leads ($-t : lag$, $+t : lead$) of the comparative series are examined for their relationship with RS_t. *VIX*, *UMCSENT* and *EPU* are compared to monthly values of RS_t, GDP growth to a quarterly

[43]The reader is encouraged to examine Figure 31 in Appendix K summarising the most representative terms of the topics.

averaged value of RS_t. The lags and leads in the first graph are therefore quarters, in the remaining they are months.

Correlations seem in general high and significant at the current period observations. For example, a correlation of more than 0.7 is found with current period nominal GDP growth. Comparing this correlation to the one of *VIX*, *UMCSENT* and *EPU* with GDP growth in Figure 15, the correlation of the Relative Sentiment Index is more substantial – in the case of *UMCSENT* more than twice as strong. RS_t appears to be correlated more strongly with *future* values of all the remaining time series. For example, there is significant correlation all the way to the third period lead of U.S. GDP growth[44]. The cross-correlation function with *VIX*, *UMCSENT* and *EPU* starts-off high (between 0.4 and 0.6) and appears to describe a cyclical relationship between RS_t and these three indicators. These findings lend further support to the notion that the Relative Sentiment Index captures actual cyclical behaviour of the economy – if RS_t is high today, it seems reasonable to expect low VIX_t (today) and high VIX_s where approximately $s > 48$, or some four years from today. The sign of the correlations also appears reasonable throughout. Recall that Relative Sentiment was constructed to increase if expansionary (contractionary) sentiment becomes more (less) pronounced. Therefore, the correlation sign is positive on GDP growth and consumer sentiment: higher growth and a more optimistic consumer sentiment go with a more expansionary/less contractionary sentiment as based on RS_t. The contrary goes for *VIX* and *EPU*, which increase with *more* uncertainty in the economy, and thus the correlation sign should be expected negative.

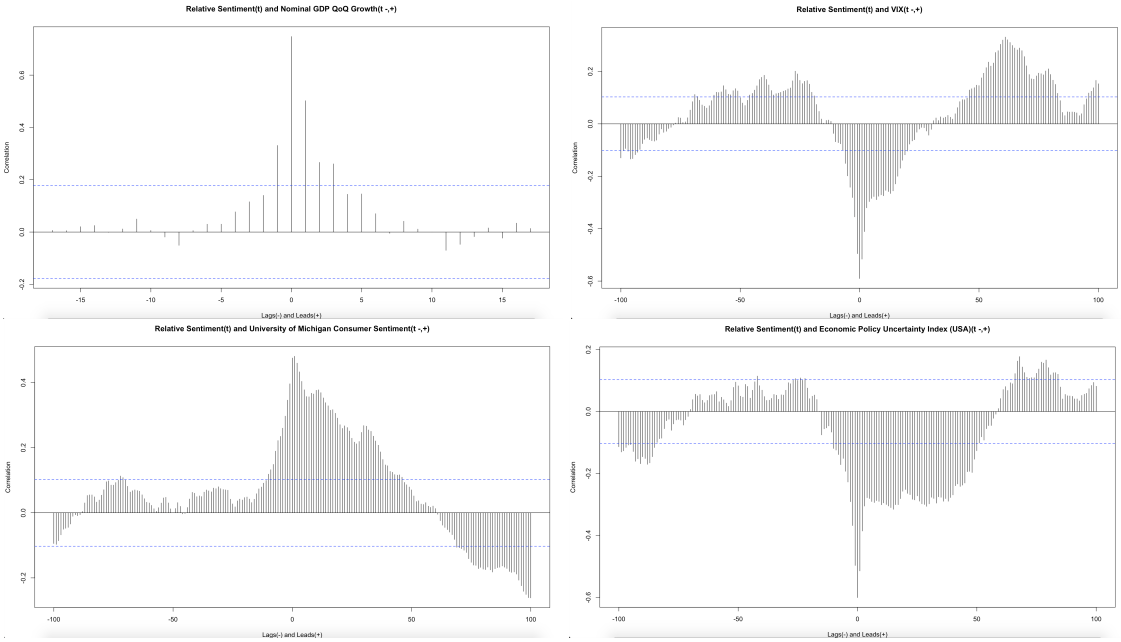


Figure 14: Cross-Correlation Functions. RS_t correlated with $Comparison_s$ where $s = (-15, ..., 15)$ in the first graph and $s = (-100, ..., 100)$ in graphs 2-4. *Comparison* refers to the four comparable time series: Nominal GDP quarter on quarter U.S.growth, CBOE Volatility Index, University of Michigan Consumer Sentiment Index and the Economic Policy Uncertainty Index. Correlation with $s > 1$ should be suggestive of predictive power. Appendix H.2 features a brief explanation of cross-correlation functions.

Comparing the relationship between RS_t and GDP growth with that of the remaining indicators and GDP growth is too worth a mention. Out of all indices, RS_t seems to be most 'visionary', predictive one. Where *UMCSENT* appears to be merely backwards-looking (significant only with lags of growth), *VIX* is merely significant at the first lag and lead. *EPU* is interesting – it seems to be somewhat significant with leads of GDP growth too, although more weakly than RS_t, and is also weakly significant with several lags of GDP growth. Moreover, the cross-correlation function of RS_t appears to be the one best capturing the purported epidemiological pattern of narratives postulated by Shiller (2017) and Shiller (2019). It rises quickly and subsequently decreases more slowly on the lead side.

[44]Under the best performing lexicon, this significant correlation was found up to the fifth lead of quarterly GDP growth.

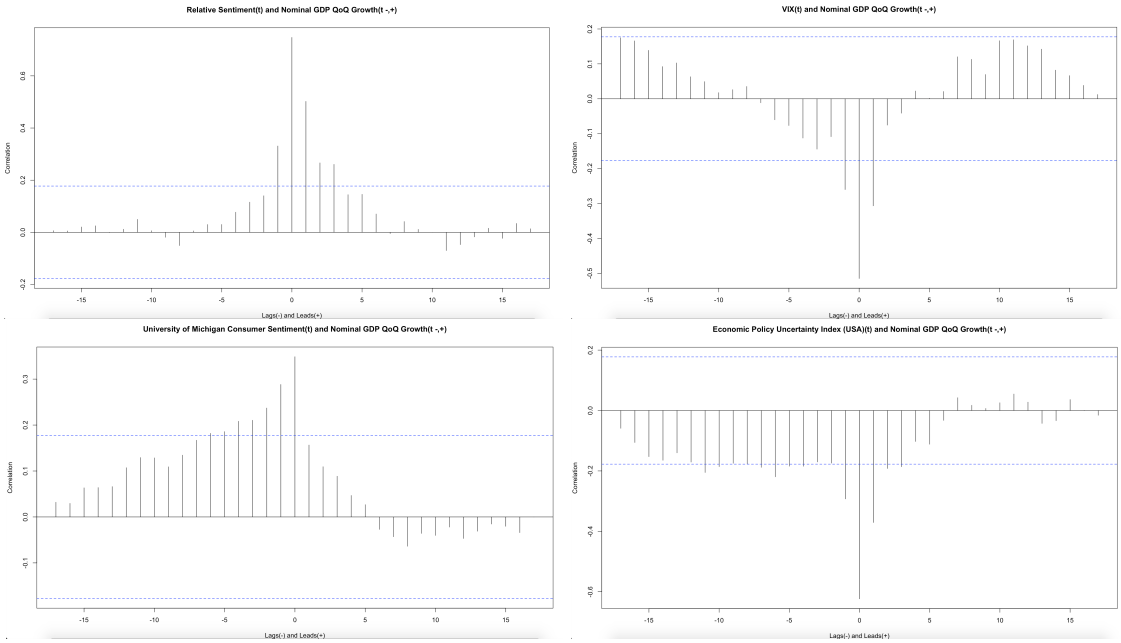


Figure 15: Cross-Correlation Functions. RS_t and $Comparison_t$ correlated with Nominal-GDP-QoQ-US-Growth$_s$ where $s = (-15, ..., 15)$ and *Comparison* refers to three comparable time series: CBOE Volatility Index, University of Michigan Consumer Sentiment Index and the Economic Policy Uncertainty Index. Correlation with $s > 1$ should be suggestive of predictive power. Appendix H.2 features a brief explanation of cross-correlation functions.

5.3.2 Regressions and In-the-Sample and Out-of-Sample Predictions

The next step in the evaluation of the predictive power of the indices towards the macroeconomy is the regression analysis. The reason for looking at the index in this simple framework is straightforward – we can gauge whether the text-based time series have any chance *at all* to predict the evolution of GDP growth. Linear regression is a straightforward model, and as such, decreases our chances of identifying a predictive relationship. To the best of my knowledge, this is the first attempt to predict, instead of merely nowcast, GDP growth with purely text-based input in the economic literature. The method is kept as simple as possible, constructing simple linear regressions between lags of the constructed text-based indices and the outcomes of GDP growth in the U.S. Subsequently, the predictive success will be evaluated in an in- and out-of-sample framework, examining the correctness of the sign, trend predictions and level errors.

The simple linear regression framework used suffers from potentially substantial omitted variable bias as well as (co-)integration issues. Being fully aware of these shortcomings, note that the purpose of the analysis is not to find a causal relationship, it is merely to get consistent estimates of the regression coefficients. For this reason, cointegration between the indices is examined. The presence of cointegration should allow the argument that the relationship found is *not* spurious (cf. cointegration testing in Appendix I.3). The econometric framework used here is elaborated on in Appendix H.3.

Kindly observe the results from four different linear regression models in Table 4. First two are regressions of GDP growth on the first lag of the Relative Sentiment Index (1), and on the first lag of growth with one lag of Relative Sentiment Index (2). In both of these regressions, Relative Sentiment is shown to be strongly significant. The introduction of the first lag of RS_t, RS_L1, leads to the coefficient on lagged nominal GDP growth becoming insignificant. The insignificance of lags of GDP is a robust result; for any linear regression model specification that involved lagged Relative Sentiment, lags of growth were never found to be significant. Successively, further lags of RS_t are introduced in the model (3), and lags of $Entropy_t$ are introduced in the model (4). Lags of $Entropy_t$ seem to be almost, but not significant when lags of RS_t are also included. This finding is slightly

Table 4: Evaluation: Linear Regression – Modelling Next Period GDP Growth

	Dependent variable:			
	Nominal_GDP_QoQ_Growth			
	(1)	(2)	(3)	(4)
Nominal_GDP_QoQ_Growth_L1		0.198		
		(0.169)		
RS_L1	668.883***	570.782***	787.537***	803.329***
	(87.497)	(120.994)	(122.540)	(124.549)
RS_L2			−296.845*	−279.371*
			(163.978)	(164.713)
RS_L3			454.097***	480.367***
			(172.138)	(171.994)
RS_L4			−398.904**	−440.299**
			(169.911)	(170.438)
RS_L5			269.014	290.579*
			(168.888)	(169.806)
RS_L6			−146.413	−188.622
			(135.897)	(137.148)
Entropy_L1				4.106
				(2.889)
Entropy_L2				−5.864
				(3.715)
Entropy_L3				5.651
				(3.692)
Entropy_L4				−4.759
				(2.880)
Constant	4.163***	3.270***	4.098***	7.252
	(0.323)	(0.827)	(0.326)	(6.503)
Observations	121	121	116	116
R^2	0.329	0.337	0.394	0.423
Adjusted R^2	0.324	0.326	0.361	0.368
Residual Std. Error	3.552 (df = 119)	3.546 (df = 118)	3.498 (df = 109)	3.478 (df = 105)
F Statistic	58.440*** (df = 1; 119)	29.998*** (df = 2; 118)	11.834*** (df = 6; 109)	7.710*** (df = 10; 105)
Note:				*p<0.1; **p<0.05; ***p<0.01

unfortunate; the expectation was to find a strong relationship[45]. Inclusion of Entropy removes the effect of the intercept in the model (4). The F-statistic of all models is consistently significant, and the adjusted R squared varies from approximately 0.32 to 0.37. The improvement with the inclusion of more lags of both indices is only marginal. Under experimenting with hyperparameters in constructing both indices, slightly higher R squared values sometimes emerged. The ones reported here are rather low but note that we have solely used text to explain future GDP growth outcomes.

Performing actual forecasts with the model, current period, one period ahead, and two-period ahead GDP growth forecasts are attempted, based on specifications as in (4). For the current period nowcast, current values of RS_t and $Entropy_t$ are added to the regression. For the two periods ahead prediction, RS_L1 and $Entropy\_L1$ are scrapped. These regressions are firstly performed on the entire series (1990-2020 observations), and in-the-sample forecasts are generated. The results of this are presented in Table 5. Thereafter, the model is estimated on the first 79 observations, and the remaining 36 observations are forecasted out-of-sample. These results are presented in Table 6. The correctness of the sign, as well as of the level ($< 2\%$ and $< 1\%$ error), is examined. Furthermore,

[45]The reader should nevertheless note, that Entropy was found to be significant when a smaller number of topics is chosen to estimate the LDA model, in interaction terms, or in different combinations with lags of Relative Sentiment. Further results can be found in Appendix M, Table 19. Interaction of Entropy and Sentiment is an interesting case study on its own, and could be the essence of the third proposition by Shiller (2019) outlined in Section 2.2.

Table 5: Evaluation: Linear Regression Predictions – In-the-Sample Forecasts

Nominal GDP QoQ Growth: Predicting Current Quarter					RMSE: 2.65
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	110	74	87	80	42
Incorrect	5	41	28	35	73
% Correct	95.7%	64.3%	75.7%	69.6%	36.5%
Nominal GDP QoQ Growth: Predicting Next Quarter					RMSE: 3.31
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	109	63	83	68	35
Incorrect	6	52	32	47	80
% Correct	94.8%	54.8%	72.2%	59.1%	30.4%
Nominal GDP QoQ Growth: Predicting Two Quarters Ahead					RMSE: 3.91
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	109	57	78	70	39
Incorrect	6	58	37	45	76
% Correct	94.8%	49.6%	67.8%	60.9%	33.9%

Note: The figures reported are from in-the-sample forecasts based on specifications as in Model (4) in Table 4 where current period regressors were added for current period forecast and first lag regressors were scrapped for two-period ahead forecast. The first column refers to whether the forecast had the correct sign (as compared with the actual realisation of the GDP growth). The second and third refer to the forecast of a trend – whether GDP growth will increase or decrease in the predicted period – if predicted based on last forecast or current period GDP growth value. The last two columns give the proportion of forecasts that had at least somewhat correct level – smaller than 2 % and 1 % deviation.

Table 6: Evaluation: Linear Regression Predictions – Out-of-Sample Forecasts

Nominal GDP QoQ Growth: Predicting Current Quarter					RMSE: 1.63
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	36	24	27	23	13
Incorrect	0	12	9	13	23
% Correct	100%	66.7%	75.0%	63.9%	36.1%
Nominal GDP QoQ Growth: Predicting Next Quarter					RMSE: 2.17
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	35	19	24	22	17
Incorrect	1	17	12	14	19
% Correct	97.2%	52.8%	66.7%	61.1%	47.2%
Nominal GDP QoQ Growth: Predicting Two Quarters Ahead					RMSE: 2.31
Prediction Type	Sign	Trend (Base: Last Forecast GDP)	Trend (Base: Last True GDP)	$< 2\%$ Error	$< 1\%$ Error
Correct	34	19	28	26	16
Incorrect	2	17	8	10	20
% Correct	94.4%	52.8%	77.8%	72.2%	44.4%

Note: Trained on the first 79 observations. Predictions generated for 43 pseudo-out-of-sample observations. The figures reported are from these out-of-sample forecasts based on specifications as in Model (4) in Table 4 where current period regressors were added for current period forecast and first lag regressors were scrapped for two-period ahead forecast. The first column refers to whether the forecast had the correct sign (as compared with the actual realisation of the GDP growth). The second and third refer to the forecast of a trend – whether GDP growth will increase or decrease in the predicted period – if predicted based on last forecast or current period GDP growth value. The last two columns give the proportion of forecasts that had at least somewhat correct level – smaller than 2 % and 1 % deviation.

it is looked at whether the trend was correctly predicted – whether the forecasts correctly predicted an increase or a decrease in GDP growth – which is performed for both the last true available GDP growth value and the last forecast as reference variables. When the forecast is higher (lower) than last available forecast (base: last forecast GDP) or last true GDP (base: last true GDP), the 'up' ('down') trend is noted. Then it is evaluated against the actual, true outcome.

The reader will note that most of the predictions have large incorrectness margins, perhaps with the notable exception of the sign and trend (base: last true GDP) predictions. Remarkably, the predictions do not get unduly worse the longer forward one estimates. Particularly, if one examines the quality of the out-of-sample predictions Table 6, the accuracy was found similar, if not better, than of the in-the-sample forecasts. That could suggest

that the long-term relationship between the text-based indices and GDP growth is somewhat stable.

As a robustness check, we might wish to examine the regression 'turned-around' in Table 7 where RS_t is regressed on lags of GDP growth – models (1) and (2), and on the lags of RS_t – models (3) and (4). GDP growth is only ever significant with one lag, and this relationship disappears once further lags of RS_t are added in models (3) and (4). Furthermore, it was examined whether the Relative Sentiment Index could be cointegrated with GDP growth. The reason for this is that quarter on quarter GDP U.S. growth, both nominal and real, were found to be integrated of order 1. The same is true for RS_t but not for $Entropy_t$ (cf. Tables in 16) *in quarterly intervals*. A series of augmented Dickey-Fuller tests were performed (cf. Table 15 and 16 in Appendix I.1). For the monthly series, both RS_t and $Entropy_t$ are found stationary. As can be seen from Table 18 in Appendix I, there is ample evidence of cointegration between RS_t and GDP growth series. Therefore, most of the regressions above should not be spurious, and the estimated coefficients consistent. There seems to exist a long-term relationship between Relative Sentiment and GDP growth.

Table 7: Robustness: Modelling Next Period Relative Sentiment Index with GDP Growth

	Dependent variable:			
	RS			
	(1)	(2)	(3)	(4)
RS_L1			0.825***	0.806***
			(0.105)	(0.124)
RS_L2				−0.031
				(0.148)
RS_L3				0.177
				(0.147)
RS_L4				−0.078
				(0.125)
Nominal_GDP_QoQ_Growth_L1	0.001***	0.001***	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Nominal_GDP_QoQ_Growth_L2		0.0001	−0.0002*	−0.0002
		(0.0001)	(0.0001)	(0.0001)
Nominal_GDP_QoQ_Growth_L3		0.0001	0.0001	0.00000
		(0.0002)	(0.0001)	(0.0001)
Nominal_GDP_QoQ_Growth_L4		−0.00001	−0.00003	−0.00004
		(0.0001)	(0.0001)	(0.0001)
Constant	−0.004***	−0.004***	0.00004	0.0004
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	121	118	118	118
R^2	0.294	0.321	0.563	0.569
Adjusted R^2	0.288	0.297	0.543	0.537
Residual Std. Error	0.004 (df = 119)	0.003 (df = 113)	0.003 (df = 112)	0.003 (df = 109)
F Statistic	49.637*** (df = 1; 119)	13.344*** (df = 4; 113)	28.836*** (df = 5; 112)	17.990*** (df = 8; 109)
Note:				*p<0.1; **p<0.05; ***p<0.01

5.3.3 Granger Causality

To further gauge the nature of the relationship between the text-based indices and U.S. GDP growth as well as other similar macroeconomic indicators, a variety of Granger causality tests were performed. Granger causality, at its core, is not about causality per se but is asking the question whether the information contained in one time series could help to predict outcomes in another series. Ideally, we would hope that there is evidence in favour of the indices Granger-causing the macroeconomic indicators and at the same time *no* evidence of this in the opposite direction – the comparative series Granger-causing RS_t and $Entropy_t$.

Remarkably, this appears to be the case for the relationship between RS_t and nominal GDP growth. The Wald

statistic from comparing the bivariate model (including RS_t) when predicting GDP growth to the univariate model (only lags of GDP growth included) is strongly significant. Kindly observe Table 8 for the result. At the same time, the statistic testing whether GDP growth has predictive power towards the RS_t index – in Table 9 – is found insignificant. For real GDP growth, the same holds as well, although real GDP growth was found to Granger-cause RS_t at the weakest, 10% confidence interval too. The relationships of RS_t with *VIX* and *EPU* are too worth mentioning. There appears to be evidence of Granger causality in both directions, which can perhaps be interpreted as evidence that, similar as these indices are to RS_t, they all capture *different*, but related aspects of the macroeconomic reality. *UMCSENT* was also found to be predictable with the Relative Sentiment series.

Table 8: Granger Causality: Constructed Series → Comparative Series

Granger Causality: Constructed Series → Comparative Series					
Series	GDP, nom, QoQ, growth	GDP, real, QoQ, growth	VIX	UMCSENT	EPU
Relative Sentiment	19.8^{***} $(8.5 * 10^{-6})$	15.4^{***} $(8.9 * 10^{-5})$	13.3^{*} (0.065)	30.9^{***} $(2.6e-05)$	32.4^{***} (0.0002)
Entropy	0.41 (0.52)	0.2 (0.65)	0.39 (0.94)	0.23 (0.89)	8.3^{*} (0.08)

Note: This table reports the results of Granger causality tests according to the approach by Toda and Yamamoto (1995). The null hypothesis is that of no Granger causality. The question here is whether RS_t and $Entropy_t$ Granger-cause the comparative time series. The statistic reported is the Wald statistic having Chi-squared distribution determined by degrees of freedom given by order of the VAR chosen with AIC. Numbers in parentheses are the corresponding p-values of the test. See Appendix H.4 for further details on how the testing was performed. *p<0.1; **p<0.05; ***p<0.01

Looking at $Entropy_t$, the relationships are much less pronounced. There does not appear to be evidence of Granger causality in any direction – with the notable exception of *EPU*. This outcome is again somewhat surprising, given the previous reasoning and expectations, and the fact that in some specifications of the linear regressions the Entropy Index was (weakly) significant, particularly in interaction terms (cf. Table 19). It is, however, difficult to hypothesise about potential reasons. The predictive power of Entropy for *EPU* is perhaps not surprising. Remember that *EPU* is based on counting news articles that include policy uncertainty terms. It could then be expected that narrative consensus, as the broadness of the dialogue, relates to article counts in general. Especially, one could argue that a broader dialogue prerequisites a larger amount of news articles. There is nothing in the LDA model to say that this *must* be true, but it is a likely presumption and a possible explanation why these two indices were found related while $Entropy_t$ looks to have no predictive power for any other time series.

Table 9: Granger Causality: Comparative Series → Constructed Series

Granger Causality: Comparative Series → Constructed Series		
Time Series	Relative Sentiment	Entropy
GDP, nominal, QoQ Growth (U.S.), quarterly	1.2 (0.27)	0.12 (0.73)
GDP, real, QoQ Growth (U.S.), quarterly	3.2^{*} (0.075)	0.003 (0.95)
VIX (CBOE Volatility Index), monthly average	21.5^{***} (0.003)	2.6 (0.46)
University of Michigan Consumer Sentiment (UMCSENT)	8.4 (0.21)	0.14 (0.93)
Economic Policy Uncertainty (EPU), U.S., monthly	30.5^{***} (0.0004)	1.7 (0.8)

Note: This table reports the results of Granger causality tests according to the approach by Toda and Yamamoto (1995). The null hypothesis is that of no Granger causality. The question here is whether the comparable time series Granger-cause RS_t and $Entropy_t$. The statistic reported is the Wald statistic having Chi-squared distribution determined by degrees of freedom given by order of the VAR chosen with AIC. Numbers in parentheses are the corresponding p-values of the test. See Appendix H.4 for further details on how the testing was performed. *p<0.1; **p<0.05; ***p<0.01

In summary, observing the results in these past sections, there appears to be ample evidence that the Relative Sentiment Index has predictive power when it comes to GDP and several other uncertainty- and sentiment-capturing macroeconomic indicators. The evidence on the predictive power of the Entropy Index is rather weak.

5.3.4 Structural Break Analysis

Finally, we examine whether substantial changes in the value of the text-based indices, particularly the Relative Sentiment, could potentially serve as early-warning crisis indicators. The method of Bai and Perron (2003) was

used to partition the indices by structural breaks of their intercept – the mean. The approach estimates the *optimal* number of partitions based on the BIC criterion. It was found that whether the multiple structural break approximation is performed on the quarterly aggregated RS_t series or the original, monthly series, *every* official – National Bureau of Economic Research (n.d.) defined (peak to through) – crisis since 1990 is identified in advance, with some lead over the actual onset of the crisis. In the figures below, the reader can see the structural breaks resulting from the optimising approach of Bai and Perron (2003)[46]. The quarterly RS_t series performs remarkably well. Appendix N, Figure 33, features the corresponding plots for $Entropy_t$. Figure 34 plots the BIC statistics for a varying number of breaks. The econometric framework applied here is elaborated in Appendix H.5.

Table 10: U.S. Official Business Cycle Turning Points: Dates by National Bureau of Economic Research (n.d.)

BUSINESS CYCLE REFERENCE DATES		DURATION IN MONTHS			
Peak	**Trough**	**Contraction**	**Expansion**	**Cycle**	
Quarterly dates are in parentheses		Peak to Trough	Previous trough to this peak	Trough from Previous Trough	Peak from Previous Peak
July 1990(III)	March 1991(I)	8	92	100	108
March 2001(I)	November 2001 (IV)	8	120	128	128
December 2007 (IV)	June 2009 (II)	18	73	91	81
February 2020 (2019 IV)	–	–	128	–	146

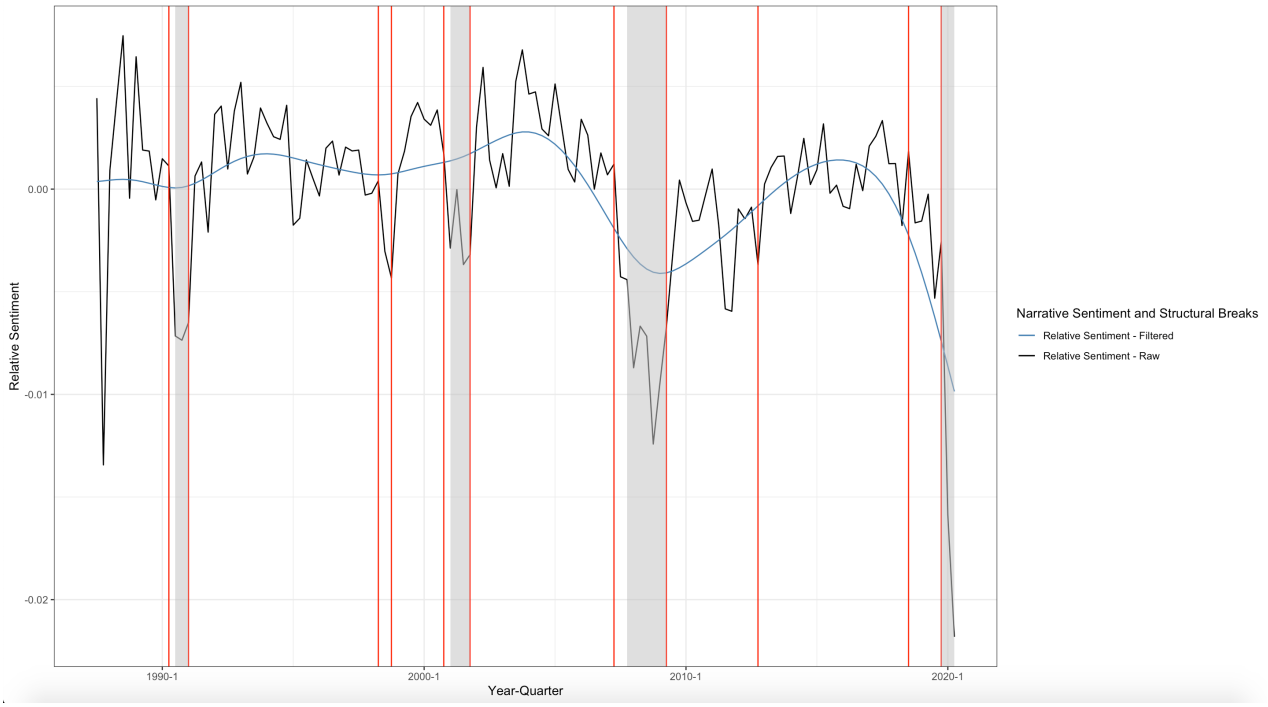


Figure 16: Quarterly Estimated Optimal Structural Breaks (BIC Chosen Model) – Red Colour

The breaks in monthly RS_t Index are found to have occurred in Jun 1990, Feb 1991, Jun 1998, Jan 1999, Nov 2000, Dec 2001, Jul 2007, Sep 2008, May 2009, Feb 2013 and Jul 2019. In the quarterly series, the breaks are set in 1990 Q2, 1991 Q1, 1998 Q2, 1998 Q4, 2000 Q4, 2001 Q4, 2007 Q2, 2009 Q2, 2012 Q4, 2018 Q3, 2019 Q4.[47] As Figures 16 and 17 highlight, it is especially the financial crisis which stands out and is recognised particularly well based on the narrative sentiment analysis – with a lead of two quarters, or five months versus the official

[46]The breaks are found autonomously. The only parameter that was set exogenously is the minimum amount of periods in each partition – for quarterly series it is $h = 2$ and for the monthly $h = 6$. These are informed guesses based on NBER's dates.

[47]If the model is *forced* to look for *exactly* seven breaks – there were exactly seven breaks since 1990 as defined by National Bureau of Economic Research (n.d.) – the model identifies all breaks as NBER would – all peaks and throughs are identified either exactly or with some lead.

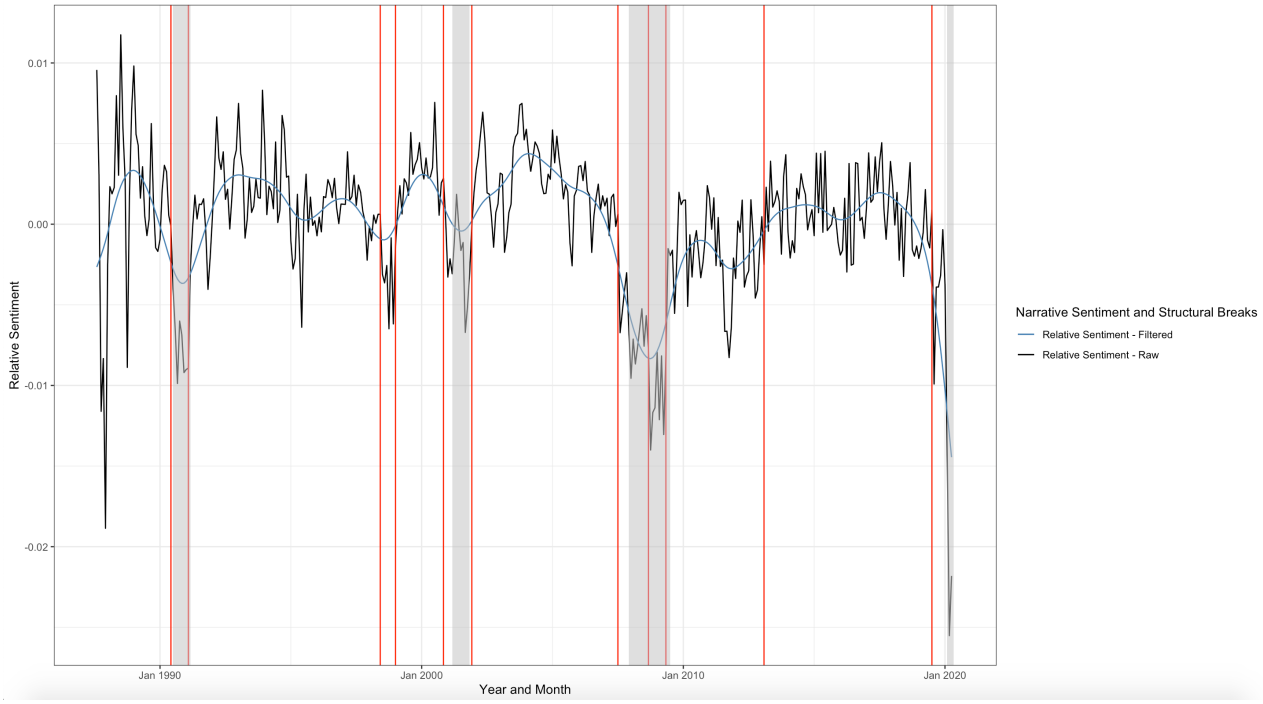


Figure 17: Monthly Optimal Estimated Structural Breaks (BIC chosen model) – Red Colour

NBER estimate. Similarly, it is the early 2000s, dot-com bubble, recession that is also identified with a lead of one quarter, or four months in the monthly series. With the latter, we can argue that a clear risk of a crisis was visible even earlier. The monthly series identifies a small 'crisis' already in the late-1990s, around the time of the aftermath of the 1997 Asian financial crisis. This interpretation seems reasonable as it was the time where the common perception of the economy shifted towards general doubts about the sustainability of the boom at that time. The identification of a crisis concluding – the 'through' in NBER's nomenclature – was found more challenging and less predictively pronounced in the breaks.

It should be noted that finding structural breaks *ex-ante* in this manner is not possible; the analysis above is purely descriptive. Nevertheless, it should be enough to argue that if a major shift in the Relative Sentiment transpires, particularly one that does *not* correct over the following period(s), one can reasonably expect that a notable macroeconomic shift has truly occurred. As can be concluded from the Figures 16, 17 and Table 10, the large shifts identified by structural break analysis have a 0% false-negative ratio in terms of (early) crisis warning. Before each crisis, there appears to have been a major correction downwards in the Relative Sentiment Index.

5.4 Robustness: Word Embeddings and Narrative Lexicons

To lend further strength to the results presented above, it makes sense to perform two more important tests. The most apparent critique towards the machine learning model presented, and how the index was constructed, is perhaps the fact that the vector space – the derived word embeddings – were created with news articles spanning the entire 30 years. It could be, therefore, pointed out that the expansionary and contractionary lexicons possess, by construction, information about the future, and thus it could not have been credibly demonstrated that there exists a correlation between news reporting content and future macroeconomic outcomes. Two further robustness checks shall address this crucial critique.

Firstly, we could examine both of the lexicons for terms which could possibly not have been identified, had we not estimated the word embeddings with the entire corpus on business cycle news. To illustrate, if there are words in

the contractionary lexicon which are terms commonly used to refer to specific crises, or are known to have been the cause of the crisis in the past as of today, they could potentially not have been identified if we had estimated the word embeddings on only an early subset of news articles. For example, the words *coronavirus* as well as *covid* can be found in the contractionary lexicon. Would the model already have found these in late 2019? When the model is run on 1990-2019 data, these words are not identified with the set cut-off of cosine similarity. Therefore, let us select these obviously 'visionary' words, or the ones that are clearly associated with a specific one crisis (or boom), and then examine whether the results change. Demonstratively, the following tokens (unigrams and bigrams) are removed from the contractionary lexicon: *severe financial*, *financial crisis*, *recession financial*, *worst financial*, *banking crisis*, *credit crisis*, *economy financial*, *global financial*, *asian crisis*, *virus*, *respiratory*, *coronavirus*, *sars*, *currency crisis*, *covid*, *debt crisis*. From the expansionary lexicon, the following two are removed: *gold's* and *buy cars*. All the remaining words are arguably highly generic, and commonly used unigrams and bigrams that could have been used at any point in the history, and thus should not be particularly relevant to a certain single crisis.

For the sake of brevity, I do not plot all the results as in Sections 5.1 and 5.3. The results, particularly on the cross-correlation and structural breaks, remain largely unchanged. See Appendix O for these results. The time series of contractionary, expansionary as well as relative sentiment show similar interesting patterns as before, albeit the indices are now more variable and somewhat noisier. In Table 11, the reader can find the results of Granger causality testing this *Robust* Relative Sentiment Index. The strength of the predictive relationship in Table 11 – denoted Robust 1 – as compared to Table 8 is found somewhat weaker, but nevertheless consistent with previous results. The only substantial difference is the insignificance on the predictive relationship for VIX and significance of Granger causality from UMCSENT to Relative Sentiment.

Table 11: Granger Causality: Relative Sentiment (Robust) → Comparative Series

Granger Causality: Relative Sentiment (Robust) → Comparative Series					
Series	GDP, nom, QoQ, growth	GDP, real, QoQ, growth	VIX	UMCSENT	EPU
Relative Sentiment (Robust 1)	6.2^{**} (0.013)	3.7^{*} (0.053)	11.0 (0.14)	21.4^{***} (0.002)	18.3^{***} (0.006)
Relative Sentiment (Robust 2)	5.6^{**} (0.018)	5.6^{**} (0.018)	2.7 (0.26)	11.6^{**} (0.041)	6.0 (0.3)

Note: This table reports the results of Granger causality tests according to the approach by Toda and Yamamoto (1995). The null hypothesis is that of no Granger causality. The question here is whether RS_t (robust) Granger-causes the comparable time series. The statistic reported is the Wald statistic having Chi-squared distribution determined by degrees of freedom given by order of the VAR chosen with AIC. Numbers in parentheses are the corresponding p-values of the test. See Appendix H.4 for further details on how the testing was performed. *p<0.1; **p<0.05; ***p<0.01

Table 12: Granger Causality: Comparative Series → Relative Sentiment (Robust)

Granger Causality: Comparative Series → Relative Sentiment (Robust)		
Time Series	Relative Sentiment (Robust 1)	Relative Sentiment (Robust 2)
GDP, nominal, QoQ Growth (U.S.), quarterly	1.9 (0.17)	0.39 (0.53)
GDP, real, QoQ Growth (U.S.), quarterly	5.2^{**} (0.023)	2.3 (0.13)
VIX (CBOE Volatility Index), monthly average	30.3^{***} ($8.2e-05$)	12.2^{***} (0.002)
University of Michigan Consumer Sentiment (UMCSENT)	14.1^{**} (0.028)	4.8 (0.45)
Economic Policy Uncertainty (EPU), U.S., monthly	28.7^{***} ($7e-05$)	16.5^{***} (0.006)

Note: This table reports the results of Granger causality tests according to the approach by Toda and Yamamoto (1995). The null hypothesis is that of no Granger causality. The question here is whether the comparable time series Granger-cause RS_t (robust). The statistic reported is the Wald statistic having Chi-squared distribution determined by degrees of freedom given by order of the VAR chosen with AIC. Numbers in parentheses are the corresponding p-values of the test. See Appendix H.4 for further details on how the testing was performed. *p<0.1; **p<0.05; ***p<0.01

In another attempt at increasing robustness of the above results, the vector space word embeddings are estimated with business cycle articles up to and including the year 2007 *only*. The index is then calculated starting January 2008 and onwards until April 2020. Granger causality is calculated for this shortened series. The results hold consistent for GDP growth, but not all remaining relationships hold. For example, the predictive relationship with EPU breaks down. See Tables 11 and 12 for the results – Robust 2 specification. Cross-correlation was found weaker but still present at leading lags with GDP growth. The series was, however, found noisier. This finding could be interpreted as further evidence of a correlation between the RS_t Index and future realisations of GDP growth. Still, the reader should note that the word embedding estimation performs worse with less data

input. A more sizable corpus of news articles would have been necessary for robustness checks such as this one. It is not clear whether the predictive power is weaker because of the (smaller) size of the news corpus or generally as a result of the method.

6 Limitations and Discussion

The next step will be to discuss limitations and future research ideas, as well as to offer insights into the learnings acquired in the process of constructing the indices. It should be valuable to discuss the insights from the above paragraphs in a more critical and prospective manner. Particularly, since many of my experiments resulted in interesting results, potential extensions and further research possibilities should be outlined.

Limitations

Firstly, let us closely examine the many problems and limits to my study. Several of these issues, such as the considerable data-related and computational limitations, were out of my immediate control. The former was expected from the start – there is only a handful of services that collect news data, and all of them limit or directly forbid text mining use of their platforms. Text mining licenses often come with a substantial financial cost, and the type of large-scale analysis attempted in the economic publications is thus inaccessible to the average student. On the other hand, the apparent data limitations forced me to think hard what news articles are relevant, and what content is not – or said differently – where the signal-to-noise ratio of text is high for the research question at hand. Therefore, I believe that the results turned out interesting as a direct result of smart ways to retrieve the most relevant news articles and because of designing clever search queries. Where Nyman et al. (2018) or Larsen and Thorsrud (2019a) have used 17 and 11 million news articles respectively – these authors analysed a *shorter* time span than this thesis – the method here used merely 31000 news articles to calculate the index and further 60000 to learn linguistic patterns about business cycles. Thus, there were often significantly less than 100 articles per month as a data source for the indices, in several cases, even less than 50. Nevertheless, results were found to be reasonable still. This goes to demonstrate that the quality of textual data is of utmost importance, and the reasoning behind constructing the corpora should not be underestimated.

Another limitation is computational power. Depending on how well the NLP models should be estimated, the researcher needs to invest significant computational time. Even at the relatively small corpus size used here, estimating one LDA model of 50 topics takes around half a day. At the same time, the estimation of a word embedding vector space necessitates 5-6 hours. These durations are perhaps not overly problematic, but they play a role if the researcher wishes to carefully evaluate each model and compare different specifications to one another. Such sensitivity analysis, tweaking hyperparameters or using different text pre-processing approaches would require estimating a far larger amount of models than has been possible in the course of the thesis writing. These additional tests could potentially be crucial to strengthen the robustness of the results and to provide reasons for why some approaches worked, and some did not. On the other hand, the literature using similar NLP tools as, for example, Larsen and Thorsrud (2019a) does not provide such hyperparameter robustness checks either.

Perhaps a surprising limiting dimension is the, arguably, only incomplete understanding of the learning processes of the NLP-algorithms utilised in this thesis in the computer science literature. The mathematical obscurity, combined with numerous assumptions, approximations, and the focus on predictive accuracy to the detriment of methodological transparency, often clouds the learning processes underlying these algorithms. Notably, it seems far from clear what information or linguistic property do word embeddings and LDA models capture and learn *exactly*. Although these methods have been demonstrated to be useful in a wide variety of tasks, there is a limited understanding of why they perform so well. The exact nature of the *latent components* which they identify remains somewhat mysterious. When it comes to the analogical relationships and distances in the vector space derived by GloVe or related algorithms, this becomes blatantly obvious. Computer science literature is currently awash with new papers scrambling to mathematically express how language semantics are captured in these condensed vectors, and what the linear relationships in such vector space tell us about the semantics of the language. It is far from obvious what the constructed indices capture at their deepest level. Therefore, I find this strand of research to be particularly interesting for the notion of economic *narratives* – we need to get better in mathematically

capturing the meaning of language. Only with a clearer mathematical framework and more precise linguistic and economic definitions can we truly analyse economic narratives.

The last point brings us to another limitation of my thesis. The literature using NLP in macroeconomic contexts is at best in its infancy. A count of *all* published papers in this area could probably be found to be in low two-digit numbers. The lack of extensive research in this field makes it challenging to provide theoretical background and reasoning for my work. There is much exploring left to do, and the method presented here is merely a first best guess at a sensible analysis. The results should provide clues about how and where to look next.

Discussion: Vision, Extensions and Further Research

As already became clear above, one limiting factor in using textual data for computational macroeconomic analysis is that there seems to be no universal economic theory to guide the researcher through the realm of *empirical* analysis of economic storytelling. A comprehensive theory that would precisely delimit what sources of text and how should be computationally analysed, is lacking. Many economists use the word *narratives* to denote many, and thoroughly different, properties of textual data. At times, I have seen narratives to mean sentiment or animal spirits (e.g., Nyman et al., 2018). In another study, narratives were used to denote macroeconomically relevant associations with a certain object or subject (e.g., Tuckett, Smith, & Nyman, 2014). Then again, I have seen the topics – the latent categories from topic modelling algorithms – be referred to as narratives (e.g., Larsen & Thorsrud, 2019a; Thorsrud, 2020). At times, the entire corpus was collectively referred to as narratives (e.g., Hollrah, Sharpe, & Sinha, 2020). Finally, Shiller (2017) likes to refer to them in terms of epidemiological attention to certain concepts and their meaning. Although each of these manifestations of what allegedly are economic narratives is related and similar to each other, they are not the same. When it comes to methods of analysis, as well as the reasons for why they matter, these can be very different. The ultimate impression is that the economic literature could focus more on the theoretical underpinnings of why stories matter, what stories, and how, so that economics could make more meaningful use of NLP and other computational techniques. Underlying this somewhat muddled impression one could have from this strand of current literature, I found there to be a stable chain of *purported* causality that is to a more or less explicit degree being assumed. I understand it as follows:

$$\text{Text} \overset{1}{\Rightarrow} \text{Stories} \overset{2}{\Rightarrow} \text{Meaning} \overset{3}{\Rightarrow} \text{Narratives} \overset{4}{\Rightarrow} \text{Beliefs} \ (\approx \text{Expectations}) \overset{5}{\Leftrightarrow} \text{Macroeconomic Outcomes}$$

Up until, and including the second step, NLP and text mining can help the researcher with its many different tools. At the latest, when the researcher tries to decide on how the meaning encoded in text relates to economic narratives, a theory is needed - perhaps an economic theory or one derived from the field of sociology. To understand how narratives can shape the process of expectation formation, possibly even a philosophical or psychological theory might be needed. Arrow 5 points to both directions, because most probably, there is cyclical feedback between the two that interacts with the narratives through arrow 4. Notably, some of the computational text analysis research deals only with nowcasting, and so is not trying to establish any power to predict the future – in such a case, arrow 5 is rather an equality than also a feedback-like implication. Still, whether it is the connection, and perhaps causality, between narratives and macroeconomic outcomes, or beliefs and macroeconomic outcomes, the literature appears not to offer much to work with when it comes to designing a method for empirical analysis of economic storytelling. The conviction narrative theory of Tuckett and Nikolic (2017) that is used to theoretically justify the text-based empirical analysis in Nyman et al. (2018) is a readable exception. For this thesis, I tried to establish a theory of why the Relative Sentiment Index bears a narrative meaning – the claim was made throughout the thesis that a high relative co-occurrence probability of words with a group of unequivocal business cycle keywords should establish that these context words are in some narrative, meaningful way related to how agents understand crises. Therefore, I believe that particularly the Relative Sentiment Index comes relatively close to be accurately able to be termed narrative-based.

Nevertheless, much more could be done in the future. Firstly, I have used a specific similarity cut-off to create the lexicon and then counted the words to calculate their proportion. Instead, the entire distribution of cosine similarities could be used. Remember that each term in the vocabulary has a cosine similarity metric towards the contractionary and expansionary narrative vectors. Therefore, each word in the corpus on economic expectations,

or a specific subset, could be assigned a weighted sentiment score. When experimenting with this approach, using cosine similarities as weights for the terms in the lexica to construct the index, the performance on the evaluation was slightly better. Secondly, the analysis could be expanded to construct the Relative Sentiment Index for the Euro-Zone and other countries as well. The reader may have noticed that I used news articles pertaining to business cycles in both Europe and the U.S. to derive the lexicons. Not focusing on a specific region should strengthen the proposition that there might be stable and universal patterns in the language of economic news reporting because if anything, the lexical basis which underlies the Relative Sentiment Index is not in any way constructed on only U.S.-related news reporting. I would expect that similar interesting results would emerge if the lexicons would have been used to construct a Euro-Zone Relative Sentiment or Entropy Index.

Furthermore, it would be interesting to amass a larger corpus. Particularly, I have found that *disregarding* headlines – the titles of news articles – when learning the word embeddings *improves* the performance of the RS_t Index. By implication, the context-learning algorithm benefits from seeing the business cycle keywords in free text instead of in only headlines. This finding makes intuitive sense, because the semantics of the language are perhaps often found in a broader context of written text, and should not be reduced to the headlines of articles. If headlines are included, by construction as given from the search queries, an enormous proportion of headlines will feature the business cycle keywords, and the context learning of the algorithm would have been very much focused on this limited segment of the articles. To obtain larger news articles corpora, potentially speaking with newspapers directly or web-scraping techniques could be employed. Lastly, a wider variety of relationships between the text-based indices and external variables could be examined. There are undoubtedly other interesting macroeconomic indicators – PMI (purchasing manager index), the stock indices (e.g., S&P 500) and so on – that could be examined for their relationship with the ones constructed here.

On the methodological side, an extension and further analysis might include pre-trained word embeddings that are available online. There were first attempts to do this, as can be seen in Shapiro et al. (2017). The use of pre-trained embeddings was not attempted here, because the inference of economic narratives wanted to be performed on specific economy-only related news reporting to be truly a domain-specific analysis. More broadly, the question begs itself as to how to utilise best the vector space spanned by the embeddings, or whether to even use the low-rank approximation of the token-token-co-occurrence matrix in the first place. One could, for example not use any of the algebraic operations (summation and subtraction) on the resulting vectors, and instead, just extract the most similar words to specific keywords, and manually identify interesting patterns and disparities. For example, most similar vectors to both the words *crisis* and *prosperity* could be found, and then only those words whose vectors are found closest to just one of these word vectors be kept for the lexicons. Yet in another approach, the researcher could try to use the statistics in the high-dimensional and sparse token-token-co-occurrence matrix without using any low-rank approximation at all. This approach would, however, lead to different problems. For example, how to account for words which never co-occur in a context window together, or which occur only a small number of times? How to account for words that co-occur closer as opposed to farther from each other in the context window? The low-rank approximation of GloVe takes care of these difficulties and offers transparent flexibility in manipulating the hyperparameters. More to the point, many other interesting NLP techniques could have been examined, for example, BERT, doc2vec or language modelling in a broader sense (such as the latest GPT-3 language model). These are all even newer algorithms than the GloVe used here and are arguably better in capturing context-based meaning – semantics – of words, sentences and documents. Regarding the Narrative Consensus Index, a potential improvement in capturing economic bubbles could be achieved if one allows the number of clusters (topics) in the LDA to vary over time, and re-defines their word distributions over time. As one can see from these paragraphs, opportunities for further exploration seem almost countless.

Finally, I believe it would be interesting to move the computational narrative analysis away from sentiments and more towards objects. The work of Tuckett et al. (2014) is a justification for this aim. Shiller (2017) also focuses on the fact that narratives thrive on objects and subjects. The focus on these characteristics of economic narratives seems both computationally and conceptually more difficult. The initial aim for the thesis was to go into this direction but was found practically impossible in the process. Of particular interest could be to construct a virality indicator that could be used to measure macroeconomic relevance of objects and subjects over time in economic discourse. Such 'trend chart' of the economy could be helpful in early warning systems, to understand the implications of narratives and beliefs for macroeconomic outcomes, but also purely as an informational tool.

7 Conclusion

In this thesis, it was attempted to develop macroeconomically relevant indices that would draw on linguistic patterns in economic news reporting to provide clues about the future of the U.S. macroeconomy. In the construction of the Relative Sentiment Index, in particular, it was postulated that recurring, identifiable patterns in contexts of business cycle keywords could be utilised to extract vocabulary which may capture the essence of fluctuations in business cycle sentiment. The key logic applied here was derived from Firth's Distributional Hypothesis, which conveys that the probability distribution over its context defines a word's meaning. Drawing on the definitions of narratives from dictionaries, and the (post-)structuralist philosophical tradition, a narrative could be seen as the element that gives a story its meaning. If meaning is what we require to capture computationally to proxy economic narratives, distributional semantics and context-based natural language processing analysis are bound to be useful in narrative economics analysis. If we then wish to identify an *expansionary* or a *contractionary* narrative, it shall be defined by the distribution over the context of precisely those, and their closely related words. In essence: The true meaning of expansion and of contraction can be found in the context of these words, and tracking the relative incidence of such meaning over time in economic news reporting could be used to predict macroeconomic outcomes. Based on a recent word vector model designed by Pennington et al. (2014), a probabilistically-underpinned description of expansionary and contractionary narrative in news reporting was constructed. These narratives were represented as collections of words in two lexicons that consisted of terms which have been found to be semantically and sentimentally related to either expansion or contraction via their relative incidence in the context of these business cycle keywords.

The second measure created – the Entropy or the Narrative Consensus Index – captured the emergence of consensus and disagreement in economic expectations news reporting. Here it was postulated that a heightened consensus in economic expectations storytelling could potentially proxy the emergence of an economic bubble, whilst the replacement of consensus by disagreement a bursting of such bubble. Tracking these developments could then give us a warning indication of the degree of sustainability of current economic expectations.

Evaluated against the initial goal of constructing a text-based measure that firstly, exhibits a stable relationship with GDP, secondly, is predictive of future GDP evolution and thirdly, is able to predict business cycle turning points, there is evidence of all three of these properties in the Relative Sentiment Index. For Narrative Consensus, the evidence is weak, if existent, especially if examined separately from Relative Sentiment. Speaking about the properties of the former, Relative Sentiment was found to be strongly correlated with a measure of U.S. quarter on quarter, seasonally adjusted, GDP growth, and with other existing macroeconomic sentiment capturing indices such as the Economic Policy Uncertainty Index, University of Michigan Consumer Confidence Survey and the CBOE Volatility Index. Secondly, this significant correlation between Relative Sentiment and GDP was found to exist with leads of GDP growth as well, the index was found to Granger-cause GDP growth and helped to explain future GDP growth realisations in a simple regression model. Finally, in combination with structural break analysis, the results seemed to provide substantial clues that economic storytelling matters, not only in terms of representing the present state of the macroeconomy, as was already amply argued in related literature, but also for predicting its outcomes in the, at least immediate, future. The correlation between present economic storytelling and future macroeconomic outcomes was established. In reference to the anecdote of beauty contest of Keynes (1936) and the financial instability hypothesis of Minsky (1992) from the introduction, the results seem to strengthen the evidence that it should be realistic to capture the business cycle 'guessing game' and key business cycle 'moments' with textual analysis.

Secondly, employing topic modelling and using the entropy metric intuitively seems a viable tool to capture the emergence of economic bubbles, but was found less attractive than anticipated in the light of the econometric evidence. The reason for this could lie in the time-invariability of the underlying topic structure (contrary to Nyman et al. (2018)) and omission of tone-adjustment in constructing an index based on topics (as done in Larsen and Thorsrud (2019a)). Albeit the Narrative Consensus Index was found econometrically uninteresting in the evaluation analysis, its graphical form exhibits intriguing patterns. The index reached its global minimum shortly before the October 1997 mini-crash and a pronounced local minimum shortly before the recent financial crisis. No such pronounced minimum exists before the current Covid-crisis. Despite the theoretical argumentation of Gennaioli and Shleifer (2020) and empirical evidence of Nyman et al. (2018) and Larsen and Thorsrud (2019a),

it remains non-trivial to capture the relationship between the alignment of economic expectations and macroeconomic outcomes in textual data. The analysis performed was not able to convincingly show that an alignment in expectations preceded major economic crises – at least not if the alignment is defined as focusing on a smaller subset of economic topics in news reporting.

Much remains to be done in the future. The broader aim of the thesis was to demonstrate the usefulness of context-based unsupervised natural language processing for analysing the macroeconomically relevant storytelling, the economic narratives as Shiller (2017) understands them. As of today, to the best of my knowledge, no published paper in economics has used context-based natural language processing techniques. I hope to have credibly demonstrated that such an avenue for further research should prove fruitful. Context is a valuable source of information in textual data and is something that the commonly utilised bag-of-words NLP techniques cannot account for. Numerous newer algorithms focusing on context learning are being developed, and the area seems to be much-promising in computer science research as well. Economists should pay attention to these developments and regularly evaluate the usefulness of these techniques for macroeconomic analyses.

In conclusion, the thesis found evidence in favour of answering the question in its title to the positive. Yes, the data provide clear clues that it could potentially be viable to predict business cycles with natural language processing – *but only to a limited degree*. Limited, certainly regarding how far ahead such a forecast can be made, and with what degree of noise and certainty.

References

Aggarwal, C. C. (2018). *Machine learning for text* (1st ed.). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-73531-3_14 doi: 10.1007/978-3-319-73531-3_14

Allen, C., Balazevic, I., & Hospedales, T. (2019). What the vec? Towards probabilistically grounded embeddings. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 7467 – 7477). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/8965-what-the-vec-towards-probabilistically-grounded-embeddings.pdf

Allen, C., & Hospedales, T. (2019, 09 – 15 Jun). Analogies explained: Towards understanding word embeddings. In K. Chaudhuri & R. Salakhutdinov (Eds.), (Vol. 97, pp. 223 – 231). Long Beach, California, USA: PMLR. Retrieved from http://proceedings.mlr.press/v97/allen19a.html

An, Z., Jalles, J. T., & Loungani, P. (2018). How well do economists forecast recessions? *International Finance*, *21*(2), 100-121. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/infi.12130 doi: 10.1111/infi.12130

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*(4), 821 – 856. Retrieved from http://www.jstor.org/stable/2951764

Antolín-Díaz, J., & Rubio-Ramírez, J. F. (2018, October). Narrative sign restrictions for SVARs. *American Economic Review*, *108*(10), 2802-29. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20161852 doi: 10.1257/aer.20161852

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2016). A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, *4*, 385 – 399. Retrieved from https://www.aclweb.org/anthology/Q16-1028 doi: 10.1162/tacl_a_00106

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in knowledge discovery and data mining* (pp. 391 – 402). Berlin, Heidelberg: Springer Berlin Heidelberg.

Azqueta-Gavaldón, A. (2017a). Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters*, *158*, 47 – 50. Retrieved from http://www.sciencedirect.com/science/article/pii/S0165176517302598 doi: https://doi.org/10.1016/j.econlet.2017.06.032

Azqueta-Gavaldón, A. (2017b). Financial investment and economic policy uncertainty in the UK. In *Proceedings of the 1st international conference on internet of things and machine learning.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3109761.3158380 doi: 10.1145/3109761.3158380

Azqueta-Gavaldón, A. (2020). Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem. *Physica A: Statistical Mechanics and its Applications*, *537*, 122574. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437119314736 doi: https://doi.org/10.1016/j.physa.2019.122574

Azqueta-Gavaldón, A., Hirschbühl, D., Onorante, L., & Saiz, L. (2020, January). *Economic policy uncertainty in the euro area: An unsupervised machine learning approach* (Working Paper Series No. 2359). European Central Bank. Retrieved from https://ideas.repec.org/p/ecb/ecbwps/20202359.html

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K., & Vašíček, B. (2013). Leading indicators of crisis incidence: Evidence from developed countries. *Journal of International Money and Finance*, *35*, 1 – 19. Retrieved from http://www.sciencedirect.com/science/article/pii/S0261560613000028 doi: https://doi.org/10.1016/j.jimonfin.2013.01.001

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1 – 22. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.659 doi: 10.1002/jae.659

Baker, S. R., Bloom, N., & Davis, S. J. (n.d.). *Economic Policy Uncertainty Index for United States [USEPUINDXM].* Retrieved 2020-11-10, from https://fred.stlouisfed.org/series/USEPUINDXM (FRED, Federal Reserve Bank of St. Louis)

Baker, S. R., Bloom, N., & Davis, S. J. (2016, July). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593-1636. Retrieved from https://doi.org/10.1093/qje/qjw024 doi: 10.1093/qje/qjw024

Balke, N. S., Fulmer, M., & Zhang, R. (2017). Incorporating the beige book into a quantitative index of economic activity. *Journal of Forecasting*, *36*(5), 497-514. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2450 doi: 10.1002/for.2450

Bauer, L. (1983). *English word-formation*. Cambridge University Press. doi: 10.1017/CBO9781139165846

Bholat, D., Hans, S., Santos, P., & Schonhardt-Bailey, C. (2015). *Text mining for central banks*. Centre for Central Banking Studies, Bank of England. Retrieved from https://EconPapers.repec.org/RePEc:ccb:hbooks:33

Bing, L., & Minqing, H. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (p. 168–177). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1014052.1014073 doi: 10.1145/1014052.1014073

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., & Simsek, (2020, January). *Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach* (Bank of England Staff Working Papers No. 848). Bank of England. Retrieved from https://ideas.repec.org/p/boe/boeewp/0848.html

Bordalo, P., Gennaioli, N., & Shleifer, A. (2018). Diagnostic expectations and credit cycles. *The Journal of Finance*, *73*(1), 199-227. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12586 doi: 10.1111/jofi.12586

Bouchet-Valat, M. (2019). tm.plugin.factiva: Import articles from 'factiva' using the 'tm' text mining framework [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tm.plugin.factiva (R package version 1.8)

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, *17*(31), 334 – 355. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8454.1978.tb00635.x doi: 10.1111/j.1467-8454.1978.tb00635.x

Brigden, A. (2019, February). *The economist who cried wolf?* Retrieved from https://www.fathom-consulting.com/the-economist-who-cried-wolf/

Bénabou, R., & Tirole, J. (2016, September). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141 – 64. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.30.3.141 doi: 10.1257/jep.30.3.141

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, *72*(7), 1775 – 1781. Retrieved from http://www.sciencedirect.com/science/article/pii/S092523120800372X (Advances in Machine Learning and Computational Intelligence) doi: https://doi.org/10.1016/j.neucom.2008.06.011

Chakrabarti, S. (2003). *Mining the web: Discovering knowledge from hypertext data*. San Francisco: Morgan Kaufmann. Retrieved from https://www.sciencedirect.com/book/9781558607545/mining-the-web doi: https://doi.org/10.1016/B978-1-55860-754-5.X5000-9

Chaubard, F., Mundra, R., & Socher, R. (n.d.). *Word-word co-occurrence matrix*. Retrieved from https://cs224d.stanford.edu/lecture_notes/LectureNotes1.pdf

Chicago Board Options Exchange. (n.d.). *CBOE Volatility Index: VIX [VIXCLS]*. Retrieved 2020-11-10, from https://fred.stlouisfed.org/series/VIXCLS (FRED, Federal Reserve Bank of St. Louis)

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*(s1), 2-9. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x doi: 10.1111/j.1475-4932.2012.00809.x

De Finetti, B. (2017). Conditional prevision and probability. In *Theory of probability* (pp. 113 – 152). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119286387.ch4 doi: 10.1002/9781119286387.ch4

Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, *94*, 65 – 76. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167923616301830X doi: https://doi.org/10.1016/j.dss.2016.11.001

Desagulier, G. (2018, April). *Word embeddings: the (very) basics*. Retrieved from https://corpling.hypotheses.org/495

Deveaud, R., Sanjuan, E., & Bellot, P. (2014, June). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 61 – 84. Retrieved from https://hal.archives-ouvertes.fr/hal-01002716 doi: 10.3166/DN.17.1.61-84

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427 – 431. Retrieved from http://www.jstor.org/stable/2286348

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*(4), 1057 – 1072. Retrieved from http://www.jstor.org/stable/1912517

Draghi, M. (2012, July). Verbatim of the remarks made by Mario Draghi. *Speech by Mario Draghi, President of the European Central Bank at the Global Investment Conference in London*. Retrieved from https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188 – 230. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440380105 doi: 10.1002/aris.1440380105

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 281–285). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/57167.57214 doi: 10.1145/57167.57214

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019, July). Towards understanding linear word analogies. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3253–3262). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1315 doi: 10.18653/v1/P19-1315

Factiva. (n.d.). New York, NY: Dow Jones & Reuters. Retrieved from http://global.factiva.com/

Feinerer, I., & Hornik, K. (2019). tm: Text mining package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tm (R package version 0.7-7)

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, Articles*, *25*(5), 1 – 54. Retrieved from https://www.jstatsoft.org/v025/i05 doi: 10.18637/jss.v025.i05

Ferrand, D., & Weil, M. (2001). *Homo narrativus, recherches sur la topic romanesque dans les fictions de langue français avant 1800.* Montpellier, France: Presses de l'Université Paul-Valéry.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. In *Studies in linguistic analysis* (Vol. 1952-59, p. 1-32). Oxford: The Philological Society. (Reprinted in F. R. Palmer (1968), pages 168 – 205)

Fisher, W. R. (1984). Narration as a human communication paradigm: The case of public moral argument. *Communication Monographs*, *51*(1), 1-22. Retrieved from https://doi.org/10.1080/03637758409390180 doi: 10.1080/03637758409390180

Franzosi, R. (1998). Narrative analysis – or why (and how) sociologists should be interested in narrative. *Annual Review of Sociology*, *24*(1), 517-554. Retrieved from https://doi.org/10.1146/annurev.soc.24.1.517 doi: 10.1146/annurev.soc.24.1.517

Franzosi, R. (2010). *Quantitative narrative analysis.* SAGE Publications, Inc. Retrieved 2020-11-04, from http://methods.sagepub.com/book/quantitative-narrative-analysis doi: 10.4135/9781412993883

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964 – 971. Retrieved from https://doi.org/10.1145/32206.32212 doi: 10.1145/32206.32212

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall CRC Press.

Gennaioli, N., Ma, Y., & Shleifer, A. (2015, June). Expectations and investment [Book]. In *NBER macroeconomics annual 2015, volume 30* (p. 379-431). University of Chicago Press. Retrieved from http://www.nber.org/chapters/c13589 doi: 10.1086/685965

Gennaioli, N., & Shleifer, A. (2020). *A crisis of beliefs.* Princeton University Press.

Gentzkow, M., Kelly, B., & Taddy, M. (2019, September). Text as data. *Journal of Economic Literature*, *57*(3), 535 – 74. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.20181020 doi: 10.1257/jel.20181020

Gittens, A., Achlioptas, D., & Mahoney, M. W. (2017, July). Skip-gram – zipf + uniform = vector additivity. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long*

papers) (pp. 69 – 76). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1007 doi: 10.18653/v1/P17-1007

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, *46*(6), 1303 – 1310. Retrieved from http://www.jstor.org/stable/1913830

Goodfriend, M., & King, R. (1997, January). The new neoclassical synthesis and the role of monetary policy [Book]. In *NBER macroeconomics annual 1997, volume 12* (p. 231-296). MIT Press. Retrieved from http://www.nber.org/chapters/c11040

Gould, S. J. (1994, October). *So near and yet so far.* Retrieved from https://www.nybooks.com/articles/1994/10/20/so-near-and-yet-so-far/

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424 – 438. Retrieved from http://www.jstor.org/stable/1912791

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111 – 120. Retrieved from http://www.sciencedirect.com/science/article/pii/0304407674900347 doi: https://doi.org/10.1016/0304-4076(74)90034-7

Greenwood, R., & Shleifer, A. (2014, January). Expectations of returns and expected returns. *The Review of Financial Studies*, *27*(3), 714 – 746. Retrieved from https://doi.org/10.1093/rfs/hht082 doi: 10.1093/rfs/hht082

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1 – 25. Retrieved from http://www.jstatsoft.org/v40/i03/

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1 – 30. doi: 10.18637/jss.v040.i13

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685 – 697. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167923613000651 doi: https://doi.org/10.1016/j.dss.2013.02.006

Hamilton, J. D. (1994). *Time series analysis.* Princeton, NJ: Princeton Univ. Press. Retrieved from http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+126800421&sourceid=fbw_bibsonomy

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 595 – 605). Austin, Texas: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D16-1057 doi: 10.18653/v1/D16-1057

Hansen, S. (2019). Introduction to text mining. In B. f. I. Settlements (Ed.), *The use of big data analytics and artificial intelligence in central banking* (Vol. 50). Bank for International Settlements. Retrieved from https://EconPapers.repec.org/RePEc:bis:bisifc:50-09

Hansen, S., McMahon, M., & Prat, A. (2017, October). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801 – 870. Retrieved from https://doi.org/10.1093/qje/qjx045 doi: 10.1093/qje/qjx045

Harmon, D., Lagi, M., de Aguiar, M. A. M., Chinellato, D. D., Braha, D., Epstein, I. R., & Bar-Yam, Y. (2015, July 17). Anticipating economic market crises using measures of collective panic. *PloS one*, *10*(7), e0131871-e0131871. Retrieved from https://doi.org/10.1371/journal.pone.0131871 doi: 10.1371/journal.pone.0131871

Hlavac, M. (2018). stargazer: Well-formatted regression and summary statistics tables [Computer software manual]. Bratislava, Slovakia. Retrieved from https://CRAN.R-project.org/package=stargazer (R package version 5.2.2)

Hodrick, R., & Prescott, E. (1997). Postwar u.s business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, *29*(1), 1 – 16. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0040360986&doi=10.2307%2f2953682&partnerID=40&md5=c6e8d2756e3146b5be9f4f3c7016951f doi: 10.2307/2953682

Hofmann, T. (2013, January). Probabilistic latent semantic analysis. *arXiv e-prints*, arXiv:1301.6705.

Hollrah, C. A., Sharpe, S. A., & Sinha, N. R. (2020, January). *The power of narratives in economic forecasts* (Finance and Economics Discussion Series No. 2020-001). Board of Governors of the Federal Reserve System (U.S.). Retrieved from https://ideas.repec.org/p/fip/fedgfe/2020-01.html doi: 10.17016/

FEDS.2020.001

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5 – 10. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207003001134 doi: https://doi.org/10.1016/j.ijforecast.2003.09.015

Huang, S., Niu, Z., & Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, *56*, 191 – 200. Retrieved from http://www.sciencedirect.com/science/article/pii/S0950705113003596 doi: https://doi.org/10.1016/j.knosys.2013.11.009

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *26*(3), 1 – 22. Retrieved from https://www.jstatsoft.org/article/view/v027i03

Jevons, W. S. (1878, November 01). Commercial crises and sun-spots. *Nature*, *19*(472), 33-37. Retrieved from https://doi.org/10.1038/019033d0 doi: 10.1038/019033d0

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, *59*(6), 1551 – 1580. Retrieved from http://www.jstor.org/stable/2938278

Juglar, C. (1862). *Des crises commerciales et de leur retour périodique en france, en angleterre et aux états-unis*. Paris: Guillaumin.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430 – 454. Retrieved from http://www.sciencedirect.com/science/article/pii/0010028572900163 doi: https://doi.org/10.1016/0010-0285(72)90016-3

Keynes, J. M. (1936). *The general theory of employment, interest, and money: Interest and money*. London: Macmillan.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653 – 7670. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417414003455 doi: https://doi.org/10.1016/j.eswa.2014.06.009

Koop, G. M. (2003). *Bayesian econometrics*. John Wiley & Sons Ltd.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128 – 147. Retrieved from http://www.sciencedirect.com/science/article/pii/S0950705116303872 doi: https://doi.org/10.1016/j.knosys.2016.10.003

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119282105.ch2 doi: 10.1002/9781119282105.ch2

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1), 159 – 178. Retrieved from http://www.sciencedirect.com/science/article/pii/030440769290104Y doi: https://doi.org/10.1016/0304-4076(92)90104-Y

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica*, *50*(6), 1345 – 1370. Retrieved from http://www.jstor.org/stable/1913386

Lahart, J. (2007, August). In time of tumult, obscure economist gains currency. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/SB118736585456901047

Lahiri, K., Monokroussos, G., & Zhao, Y. (2013). The yield spread puzzle and the information content of spf forecasts. *Economics Letters*, *118*(1), 219 – 221. Retrieved from http://www.sciencedirect.com/science/article/pii/S0165176512005654 doi: https://doi.org/10.1016/j.econlet.2012.10.022

Landmann, O. (2014, January). *Short-run macro after the crisis: The end of the "new" neoclassical synthesis?* (Discussion Paper Series No. 27). Department of International Economic Policy, University of Freiburg. Retrieved from https://ideas.repec.org/p/fre/wpaper/27.html

Larsen, V. H., & Thorsrud, L. A. (2019a, January). *Business cycle narratives* (CESifo Working Paper Series No. 7468). CESifo. Retrieved from https://ideas.repec.org/p/ces/ceswps/_7468.html

Larsen, V. H., & Thorsrud, L. A. (2019b). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203 – 218. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304407618302148 (Annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective") doi: https://doi.org/10.1016/j.jeconom.2018.11.013

Legrand, M. D.-P., & Hagemann, H. (2007). Business cycles in juglar and schumpeter. *The History of Economic Thought*, *49*(1), 1-18. Retrieved from https://www.jstage.jst.go.jp/article/jshet2005/49/1/49_1_1/_article/-char/en doi: 10.11498/jshet2005.49.1

Lesnoff, M., & Lancelot, R. (2012). aod: Analysis of overdispersed data [Computer software manual]. Retrieved from https://cran.r-project.org/package=aod (R package version 1.3.1)

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (pp. 2177 – 2185). Cambridge, MA, USA: MIT Press.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211 – 225. Retrieved from https://www.aclweb.org/anthology/Q15-1016 doi: 10.1162/tacl_a_00134

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data.* Berlin, Heidelberg: Springer-Verlag. doi: 10.1007/978-3-540-37882-2

Lucas, R. E. (1972). Expectations and the neutrality of money. *Journal of Economic Theory*, *4*(2), 103 – 124. Retrieved from http://www.sciencedirect.com/science/article/pii/0022053172901421 doi: https://doi.org/10.1016/0022-0531(72)90142-1

Lucas, R. E. (1977). Understanding business cycles. *Carnegie-Rochester Conference Series on Public Policy*, *5*, 7 – 29. Retrieved from http://www.sciencedirect.com/science/article/pii/0167223177900021 doi: https://doi.org/10.1016/0167-2231(77)90002-1

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA, USA: MIT Press.

Merriam-Webster. (n.d.-a). Boom-and-bust. In *Merriam-webster.com dictionary.* Retrieved 2020-11-05, from https://www.merriam-webster.com/dictionary/boom-and-bust

Merriam-Webster. (n.d.-b). Downturn. In *Merriam-webster.com dictionary.* Retrieved 2020-11-05, from https://www.merriam-webster.com/thesaurus/downturn

Merriam-Webster. (n.d.-c). *How many words are there in english?* Retrieved 2020-11-03, from https://www.merriam-webster.com/help/faq-how-many-english-words

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176 – 182. Retrieved from https://science.sciencemag.org/content/331/6014/176 doi: 10.1126/science.1199644

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January). Efficient estimation of word representations in vector space. *arXiv e-prints*, arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October). Distributed representations of words and phrases and their compositionality. *arXiv e-prints*, arXiv:1310.4546.

Minsky, H. P. (1992). *The financial instability hypothesis* (Working Paper No. 74). Annandale-on-Hudson, NY. Retrieved from http://hdl.handle.net/10419/186760

Mohammad, S. M., & Turney, P. D. (2013, August). Crowdsourcing a word-emotion association lexicon. *arXiv e-prints*, arXiv:1308.6297.

Murzintcev, N. (2020). ldatuning: Tuning of the latent dirichlet allocation models parameters [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ldatuning (R package version 1.0.2)

National Bureau of Economic Research. (n.d.). *US business cycle expansions and contractions.* Retrieved 2020-11-10, from https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions (NBER, National Bureau of Economic Research, Inc.)

Nimark, K. P., & Pitschner, S. (2019). News media and delegated information choice. *Journal of Economic Theory*, *181*, 160 – 196. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022053119300110 doi: https://doi.org/10.1016/j.jet.2019.02.001

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018, January). *News and narratives in financial systems: exploiting big data for systemic risk assessment* (Bank of England Staff Working Paper No. 704). Bank of England. Retrieved from https://ideas.repec.org/p/boe/boeewp/0704.html

Oxford English Dictionary. (2003). narrative, n. In *Oed online* (3rd ed.). Oxford University Press. Retrieved 2020-11-03, from https://www.oed.com/view/Entry/125146

Palmer, D. (2007). *Structuralism and poststructuralism for beginners*. New York, NY : Writers and Readers Publishing, Inc.

Palmer, F. R. (Ed.). (1968). *Selected papers of J.R. Firth, 1952-59*. London: Longmans.

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532 – 1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D14-1162 doi: 10.3115/v1/D14-1162

Pfaff, B. (2008a). *Analysis of integrated and cointegrated time series with R* (2nd ed.). New York: Springer. Retrieved from http://www.pfaffikus.de

Pfaff, B. (2008b). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4). Retrieved from http://www.jstatsoft.org/v27/i04/

Phillips, P. C. B., & Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, *58*(1), 165 – 193. Retrieved from http://www.jstor.org/stable/2938339

Qiu, D. (2015). atsa: Alternative time series analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=aTSA (R package version 3.1.2)

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ramey, V. A., & Shapiro, M. D. (1998). Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy*, *48*, 145 – 194. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167223198000207 doi: https://doi.org/10.1016/S0167-2231(98)00020-7

Årup Nielsen, F. (2011, March). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv e-prints*, arXiv:1103.2903.

Romer, C. D., & Romer, D. H. (1989). Does monetary policy matter? A new test in the spirit of friedman and schwartz. *NBER Macroeconomics Annual*, *4*, 121-170. Retrieved from https://doi.org/10.1086/654103 doi: 10.1086/654103

Romer, C. D., & Romer, D. H. (2004, September). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, *94*(4), 1055 – 1084. Retrieved from https://www.aeaweb.org/articles?id=10.1257/0002828042002651 doi: 10.1257/0002828042002651

Romer, C. D., & Romer, D. H. (2010, June). The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks. *American Economic Review*, *100*(3), 763 – 801. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.100.3.763 doi: 10.1257/aer.100.3.763

Rorty, M. C. (1922). *Some problems in current economics*. A. W. Shaw Company.

Rudebusch, G. D., & Williams, J. C. (2009). Forecasting recessions: The puzzle of the enduring power of the yield curve. *Journal of Business & Economic Statistics*, *27*(4), 492 – 503. Retrieved from https://doi.org/10.1198/jbes.2009.07213 doi: 10.1198/jbes.2009.07213

Russell, M. A., & Klassen, M. (2018). *Mining the social web: data mining facebook, twitter, linkedin, instagram, github, and more* (3rd ed.). O'Reilly Media, Inc.

Rönnqvist, S., & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing*, *264*, 57 – 70. Retrieved from http://www.sciencedirect.com/science/article/pii/S0925231217311062 (Machine learning in finance) doi: https://doi.org/10.1016/j.neucom.2016.12.110

Sartre, J.-P. (1964). *Nausea*. New York: New Directions.

Schonhardt-Bailey, C., & Bailey, A. (2013). *Deliberating american monetary policy: A textual analysis*. The MIT Press. Retrieved from http://www.jstor.org/stable/j.ctt9qf5r7

Schumpeter, J. A. (1939). *Business cycles* (Vol. 1). McGraw-Hill New York.

Schumpeter, J. A. (1954). *History of economic analysis;*. New York: Oxford University Press.

Schumpeter, J. A., & Opie, R. (1934). *The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle*. Cambridge, Mass.: Harvard University Press.

Selivanov, D., Bickel, M., & Wang, Q. (2020). text2vec: Modern text mining framework for r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=text2vec (R package version 0.6)

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2017, January). *Measuring news sentiment* (Working Paper Series No. 2017-1). Federal Reserve Bank of San Francisco. Retrieved from https://ideas.repec.org/p/fip/fedfwp/2017-01.html doi: 10.24148/wp2017-01

Shiller, R. J. (1987, November). *Investor behavior in the october 1987 stock market crash: Survey evidence* (Working Paper No. 2446). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w2446 doi: 10.3386/w2446

Shiller, R. J. (2015). *Irrational exuberance: Revised and expanded third edition* (REV - Revised, 3 ed.). Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctt1287kz5

Shiller, R. J. (2017, April). Narrative economics. *American Economic Review*, *107*(4), 967 – 1004. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.107.4.967 doi: 10.1257/aer.107.4.967

Shiller, R. J. (2019). *Narrative economics: How stories go viral and drive major economic events.* Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctvdf0jm5

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, *1*(3). Retrieved from http://dx.doi.org/10.21105/joss.00037 doi: 10.21105/joss.00037

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach.* " O'Reilly Media, Inc.".

Sims, C., Stock, J., & Watson, M. (1990). Inference in linear time series models with some unit roots. *Econometrica*, *58*(1), 113–44. Retrieved from https://EconPapers.repec.org/RePEc:ecm:emetrp:v:58:y:1990:i:1:p:113-44

Slutzky, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica*, *5*(2), 105 – 146. Retrieved from http://www.jstor.org/stable/1907241

Stock, J. H., & Watson, M. W. (2012, May). *Disentangling the channels of the 2007-2009 recession* (Working Paper No. 18094). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18094 doi: 10.3386/w18094

Strang, G. (2019). *Linear algebra and learning from data.* Wellesley, MA : Wellesley-Cambridge Press.

Sudhahar, S., De Fazio, G., Franzosi, R., & Cristianini, N. (2015). Network analysis of narrative content in large corpora. *Natural Language Engineering*, *21*(1), 81–112. doi: 10.1017/S1351324913000247

Svirin, A. (2019, Feb). *Vector addition and subtraction.* Retrieved from https://www.math24.net/vector-addition-subtraction/

Tai, I., Olson, B., & Blessner, P. (2016). Unsupervised text mining approach to early warning system. *International Journal of Computer and Information Engineering*, *10*(4), 788 – 793. Retrieved from https://publications.waset.org/vol/112

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (1st ed.). New York: Random House.

ter Ellen, S., Larsen, V. H., & Thorsrud, L. A. (2019, October). *Narrative monetary policy surprises and the media* (Working Papers No. No 06/2019). Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School. Retrieved from https://ideas.repec.org/p/bny/wpaper/0078.html

thesaurus.com. (n.d.-a). economic expansion. In *dictionary.com thesaurus.* Dictionary.com. Retrieved 2020-11-03, from https://www.thesaurus.com/browse/economic%20expansion

thesaurus.com. (n.d.-b). various. In *dictionary.com thesaurus.* Dictionary.com. Retrieved 2020-11-03, from https://www.thesaurus.com/browse

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*(2), 393-409. Retrieved from https://doi.org/10.1080/07350015.2018.1506344 doi: 10.1080/07350015.2018.1506344

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, *66*(1), 225 – 250. Retrieved from http://www.sciencedirect.com/science/article/pii/0304407694016168 doi: https://doi.org/10.1016/0304-4076(94)01616-8

Tuckett, D., & Nikolic, M. (2017). The role of conviction and narrative in decision-making under radical uncertainty. *Theory & Psychology*, *27*(4), 501 – 523. Retrieved from https://doi.org/10.1177/0959354317713158 doi: 10.1177/0959354317713158

Tuckett, D., & Nyman, R. (2017). *The relative sentiment shift series for tracking the economy.* mimeo.

Tuckett, D., Smith, R. E., & Nyman, R. (2014). Tracking phantastic objects: A computer algorithmic investigation of narrative evolution in unstructured data sources. *Social Networks*, *38*, 121 – 133. Retrieved from http://www.sciencedirect.com/science/article/pii/S037887331400015X doi: https://doi.org/10.1016/j.socnet.2014.03.001

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453 – 458. Retrieved from https://science.sciencemag.org/content/211/4481/453 doi: 10.1126/science.7455683

University of Michigan. (n.d.). *University of Michigan: Consumer Sentiment © [UMCSENT].* Retrieved 2020-11-10, from https://fred.stlouisfed.org/series/UMCSENT (FRED, Federal Reserve Bank of St. Louis)

U.S. Bureau of Economic Analysis. (n.d.-a). *Gross Domestic Product [A191RP1Q027SBEA].* Retrieved 2020-11-10, from https://fred.stlouisfed.org/series/A191RP1Q027SBEA (FRED, Federal Reserve Bank of St. Louis)

U.S. Bureau of Economic Analysis. (n.d.-b). *Real Gross Domestic Product [A191RL1Q225SBEA].* Retrieved 2020-11-10, from https://fred.stlouisfed.org/series/A191RL1Q225SBEA (FRED, Federal Reserve Bank of St. Louis)

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324 – 342. Retrieved from http://www.jstor.org/stable/2627346

Wittgenstein, L., & Anscombe, G. E. M. (1958). *Philosophical investigations* (2nd ed.). Oxford: B. Blackwell.

Woodford, M. (2009, January). Convergence in macroeconomics: Elements of the new synthesis. *American Economic Journal: Macroeconomics*, *1*(1), 267 – 79. Retrieved from https://www.aeaweb.org/articles?id=10.1257/mac.1.1.267 doi: 10.1257/mac.1.1.267

Woodford, M. (2013). Macroeconomic analysis without the rational expectations hypothesis. *Annual Review of Economics*, *5*(1), 303-346. Retrieved from https://doi.org/10.1146/annurev-economics-080511-110857 doi: 10.1146/annurev-economics-080511-110857

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *226*, 267 – 298. Retrieved from http://www.jstor.org/stable/91170

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1 – 27. doi: 10.18637/jss.v014.i06

Zeileis, A., Kleiber, C., Krämer, W., & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics Data Analysis*, *44*(1), 109 – 123. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167947303000306 (Special Issue in Honour of Stan Azen: a Birthday Celebration) doi: https://doi.org/10.1016/S0167-9473(03)00030-6

Appendix A Business Cycle Stages and N-grams

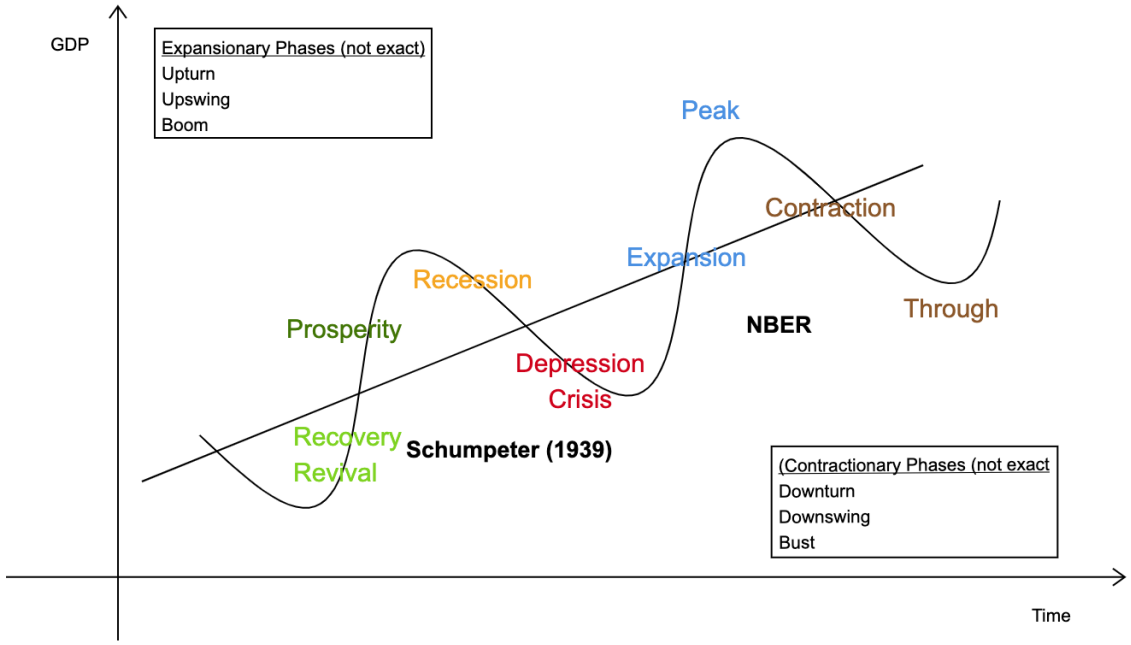


Figure 18: Stages of the Business Cycle. Based on National Bureau of Economic Research (n.d.) and Schumpeter (1939).

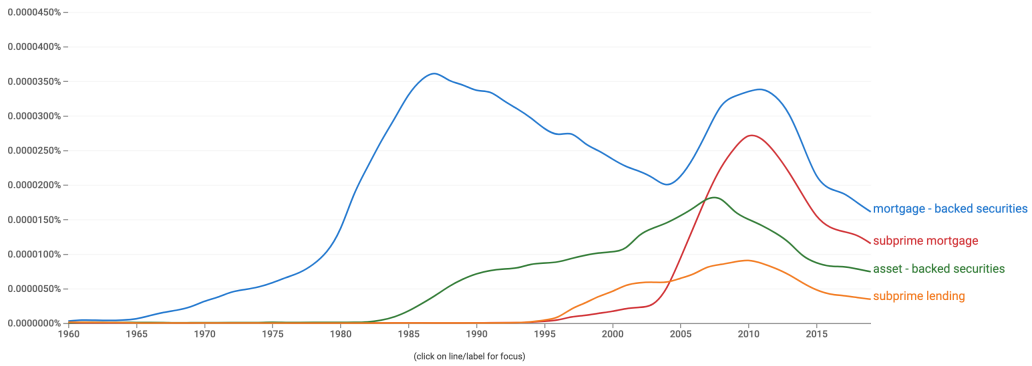


Figure 19: Mortgage-Related Products Driving the 2007-2008 Financial Crisis? Source: Google N-gram Viewer. Based on Work by Michel et al. (2011).

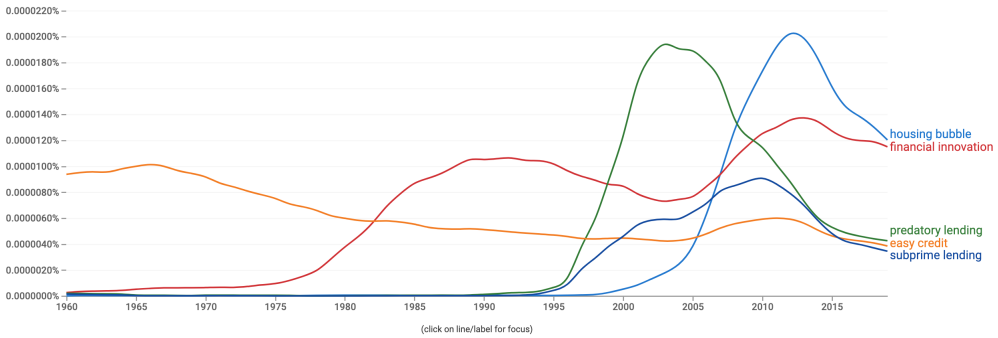


Figure 20: Other Sources of the 2007-2008 Financial Crisis? Source: Google N-gram Viewer. Based on Work by Michel et al. (2011).

Appendix B Corpus Pre-Processing

B.1 Tokenization, Stop-Words, Filtering and Text Augmentation

The first step in any textual analysis is to collect fitting data in an appropriate, machine-readable format. In Figure 2, this was referred to as pre-pre-processing and information retrieval. As *Dow Jones Factiva* is used, this step turns out to be reasonably simple. The algorithms implemented in the *Factiva* platform implement information retrieval know-how; the search queries specified in Section 3 are identified and retrieved from the platform's database via the information retrieval algorithms. Furthermore, I am using a specific *R* package, called *tm.plugin.factiva* that reads the exported *Factiva* articles into *R*, and thus pre-pre-processes automatically[48].

Having collected the raw corpora, pre-processing the data for any text mining or NLP performed afterwards is the next step. Kindly observe Figure 3 for an overview of the consecutive steps. As a first step, the package *tidyverse* in *R* is used to organise all the data in a compact format. Beyond the heading and body text information which is the main data used for textual analysis, the following information is kept as related metadata: *date stamp*, *id*, *newspaper*, *regional relevance* and *word count*. Document ids are kept, so it is always determinable from what document the textual information is coming. The next step, if desired, is to concoct text from headings and the respective body of the articles. This step shall be solely performed on the corpus on economic expectations[49]. In the next step, the data are *tokenised*. A *token* is roughly equivalent to the commonly used word *term*. The token carries semantic and syntactic meaning, particularly through the information about its context, and is, therefore, the key unit of interest for the analysis. A token does not need to be only one word. It could be a collection of words, commonly referred to as N-gram (with n being the number of words in the collection), collocation, a sentence or even an entire document. It strictly depends on the goal of our analysis, and the nature of the tools used, until what level the *tokenisation* should be performed. For this analysis, unigrams (single words) and bigrams are used. *R* has implemented a function in the package *tm* that carries out this tokenisation in an automatic manner. Punctuation is already removed by this function, as well as all words put to lower case format. This tokenisation function thus already implements parts of the filtering and augmenting steps.

After the tokenisation is completed, the corpus is examined to determine if any additional filtering – removal of tokens – is needed. To consider here is removing any remaining punctuation, the so-called stop-words, dealing with numbers (incl. in ordinal format) and removing any special, irrelevant, terms that might be present in the texts. It turns out that the R tokenisation function already does well in removing all punctuation. Tokenisation

[48] For different data formats, this procedural step could be particularly challenging. The interested reader is referred to Chakrabarti (2003) or Liu (2007) for handling HTLM-based textual data. For social media mining, Russell and Klassen (2018) is a valuable source.

[49] While learning on the corpus of business cycle news, it was found to improve results if headings were not included – only the body of the news articles was learnt upon. This could be because of the way I query the documents, the inclusion of headings would bias the algorithms to learn predominantly from the context of the headings since these per definition contain the target words of which semantics are to be learnt.

is relatively simple in English because the only information one needs to supply to the computer is that words are separated with an empty space. Hyphens are also removed, which might not always be of advantage. For example, *mortgage-backed* is separated into two words and will, therefore, only enter analysis where bigrams are present. Stop-words are, on the other hand, still largely present. These are words that often occur, albeit they have relatively little semantic relevance and are of little value for the research question[50]. A pre-specified stop-word dictionary is first used to remove these terms. *R* implements such dictionaries – the lexicon named *onix*[51] was used. It is important to realise that any words removed at this stage will not enter the later analysis and to think about potential implications. Furthermore, there is a substantial amount of numbers, especially prices, ordinal numbers or dates. These I removed as well. Finally, all 'special' terms are removed. They consist of, for example, URL links, time zone abbreviations, newspaper abbreviations, mail, newspaper tags and other special *Factiva*-database related words to only name few. These are removed with the aid of string pattern searches.

The next step is to perform any augmentations of the remaining tokens. Case-folding (reducing to lower cases) was already performed. Stemming the words is also omitted. Stemming is a method of reducing any word to its morphological root. In linguistics, the morphological root can be defined as the most basic unit of meaning which cannot be disentangled further into any morphemes[52]. The interested reader can note that there are the related, but slightly different, concepts of word *stems* and *bases* in linguistics. When computer scientists refer to stemming words, they do *not* refer to finding word stems in the linguistic sense, but they refer to finding morphological *roots*. The algorithms doing this most often either use pre-defined morphological lexicons that help the computer identify the root, or remove any suffixes, or use other heuristics while deciding on what is the root. Lemmatisation is not an error-free process. Furthermore, when *lemmatising* words, the algorithms use part-of-speech tagging and learn contexts to augment the words to its *lemma* form. Basic units of lexical meanings are called *lexemes* in linguistics, and these are groups of word stemming from a common morphological root. For example, the lexemes run, runs, and running would all be converted to run. Both the *stemming* and *lemmatisation* are omitted in the analysis here since it was found to rather confuse and distort valuable information. For example, if the adjective *secure*, and the noun (asset-backed) *security* are reduced to the same root *secur*, we lose a potentially valuable source of information. *Stemming* and *lemmatisation* are particularly important if the corpus of text used is not voluminous enough, and otherwise not enough observations of individual words would be registered. This issue, arguably, does not apply here – the corpora are substantially voluminous. Further information on text pre-processing can be found in Aggarwal (2018, p. 17-30), and on the linguistic aspects of English in Bauer (1983). Christopher Manning's and Hinrich Schütze's *Foundations of Natural Language Processing* offer a good overview of the linguistic sources of text mining and NLP (Manning & Schütze, 1999).

B.2 Data Formats, Vectors and Matrices

Both multidimensionality and sparsity of text data have been highlighted in Section 4.1 – but how do they manifest mathematically? How do we store the data? The mathematical form we use to represent textual data is highly specific for each of the purposes of further analysis. What remains constant is that the data is stored in matrices. However, the definitions of both rows and columns, and with it the nature of the matrix elements, change from algorithm to algorithm, which can sometimes obscure their workings and results. It is therefore always important to keep in mind the form of input data one feeds an algorithm.

The most commonly found matrix representation of data in NLP is the document-term matrix (DTM), or term-document matrix (TDM). Where the first dedicates each row to a document and each column to a distinct token (term), the latter uses rows for the distinct terms and columns for the documents. Suppose we name such matrix $\mathbf{A}$, its entries $A_{i,j}$ could be constructed in at least four widely used ways – zeros and ones (boolean), token

[50] Zipf's Law demonstrates how several words are over-represented in language use. It shows an inverse relationship between a word's frequency count in a corpus and its rank in the frequency distribution. To demonstrate, *the* is the most frequent word in English, with the second most frequent word being used around $\frac{1}{2}$ of the times *the* would. The third most frequent word would be expected $\frac{1}{3}$ of the times *the* would and so on. Examples of stop-words would be *the*, *a*, *that*, *and so on*.

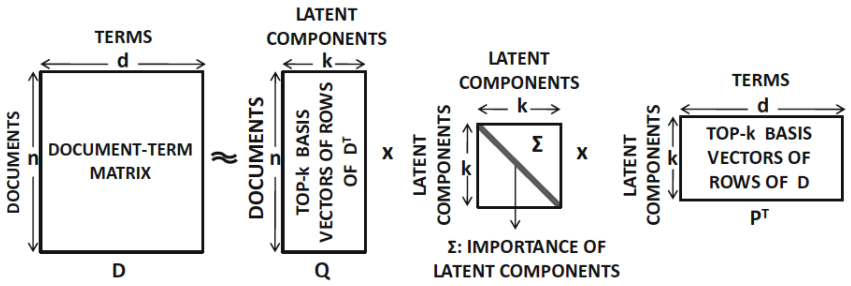
[51] Available under https://www.lextek.com/manuals/onix/stopwords.html

[52] Morphemes are mostly referred to as units of languages that carry meaning. The existence of meaning is somewhat subjective. Furthermore, a morpheme can be a word in common sense, an affix or even a combination of words.

frequencies[53], inverse document frequencies[53] or *tf-idf*'es[54]. The first method would create an entry of zero where a token is absent in the document, and a one of it is present. The second method counts the occurrences of each word in each of the texts. The third calculates how common the word is in the corpus, and the latter *tf-idf*, or term-document – inverse document frequency is proportional to the token's occurrences in a single document and its inverse document frequency, which are multiplied. For this thesis, DTM with raw word frequencies is used as an input to Latent Dirichlet Allocation.

In all of these cases, the information on the closest context of words and ordering of tokens in the documents is completely lost. This method is often referred to as *bag-of-words* models – all tokens are represented in one matrix that loses most of the structure that a text possesses. Therefore, if we want to use NLP methods which, for example, learn to understand contexts and semantic nuances, these bag of words representations would be rather disadvantageous. The most relevant for this thesis shall be the token-token-co-occurrence matrix that forms the basic building block of the word embedding algorithm. This matrix counts the number of times a *context* token (e.g., in rows) occurs in a pre-defined window (e.g., two tokens to the left and one to the right) around the main or *target* token (in columns). The algorithm utilised here uses the information in this matrix to find *relative probabilities* of co-occurrences – the ratio of probabilities of occurring next to a specific context token of two different tokens. Contrary to bag-of-words models, this data representation registers context and therefore is capable of capturing *semantics*.

Appendix C Singular Value Decomposition and Latent Structures



SVD as visualised by Aggarwal (2018, p. 37). $A = D$ and $m = d$ in the discussion below.

The strand of NLP algorithms on the discovery of latent structures began with the revival of the matrix factorisation technique called Singular Value Decomposition (SVD). Although a mathematical review of this key concept would take a textbook on itself, it is nevertheless indispensable for understanding how the computer learns these hidden structures, and why do they emerge in the first place. The use of SVD is crucial in two ways. Firstly, it combats the problem of multidimensionality and sparsity, since the factored matrices turn out compacter and less sparse with each consecutive factorisation. Secondly, it turns out that by factorising matrices, clusters representing interesting features of the textual data emerge. They emerge because text entities such as documents and words that have similar representation in this latent space tend to be related to one another, which is often used to mean that they are *similar* in some sense. These features (clusters) are then defined by a vocabulary they are found to consist of, or by a probability distribution over the vocabulary of the entire corpus. The discussion

[53]Term frequency (tf) and inverse document frequency (idf) are collective terms that include quite different approaches to the *normalisation* of the raw term or inverse document frequencies. These include different linear transformations (often natural logarithms), smoothing, or otherwise re-weigh the frequencies, so that one always needs to keep a close eye on the exact computational form of term frequency and inverse document frequency when working with these concepts in statistical software. It is not always clear what method a specific function applies without very granular inspection. It is also often not ex-ante clear, what method makes more sense for one's analysis. Manning, Raghavan, and Schütze (2008, p. 100-123) offer a good exposition of the different weightings.

[54]The tf-idf statistic is more often than in NLP used in pure information retrieval, particularly recommendation systems, as its purpose is first of all to over-weigh terms that are seldom used in a document and at the same time strongly characteristic of it. The metric estimates the degree of extraordinarity of a token in a document.

here is based on Aggarwal (2018, p. 31-71). If A is our document-term matrix, we can take its entries a perform a (full) SVD that gives

$$A = U\Sigma V^T \text{ or } A \approx U\Sigma V^T$$

where U and V^T will be orthogonal matrices (left and right singular vectors) and Σ will consist of singular values. If A is $n \times m$, we will end up with U as $n \times k$, Σ as $k \times k$ and V^T as $k \times m$. SVD is a generalisation of the concept of Eigenvectors and Eigenvalues beyond the realm of square matrices. The learning algorithms need a technique that works on matrices of any dimensionality and simultaneously reduces the dimensions. It turns out that any matrix has a unique SVD factorisation. In reducing dimensions, most algorithms do not keep the entire factorisation output but only retain some rank $r << min(m, n)$ of the matrices. This construct is then referred to as a low-rank approximation of a matrix based on the Eckart-Young Theorem. It turns out that the elements of Σ get progressively smaller from top-left to bottom-right, some of them being zero (or close to). The readers familiar with Principal Component Analysis (PCA) will recognise this feature of SVD – it is the key behind the dimensionality reduction in PCA. The zero elements (or some > 0 elements defined by a cut-off), and their corresponding multiplying columns and rows from U and V, are then removed in what is finally referred to as a reduced SVD. If a specific cut-off is used for the singular values, one refers to truncated SVD. Such approximation, however, requires an optimisation technique since the left and right singular vectors would change. This optimisation procedure is the essence of machine learning here. The low-rank approximation of the initial matrix compresses the most important information into only several latent components that are often able to capture interesting linguistic patterns[55]. Usually, this is achieved in a manner resembling least square minimisation with a concept from linear algebra called *Frobenius Norm*. One could write this optimisation problem as

$$\min_{U,V} ||A - U\Sigma V^T||_F^2 \;\; s.t. \;\; U, V^T \text{ being orthogonal}$$

Appendix D Intuition: Latent Dirichlet Allocation

Why should we search for probabilistic mixture models of text? Most prominently, there is a theoretical reason based on the work of the mathematician De Finetti. Based on De Finetti (2017), a theorem named after him emerged which dictates that for any set of *exchangeable* random variables, there exists a (latent) parameter, conditional on which the observed (random) variables are identically and independently distributed. The latent parameter enabling the conditional independence then establishes the mixture components (= topics) of the mixture model. Notice how the assumption of the *bag-of-words* model embodies the nature of statistical exchangeability – one disregards the order of the tokens in the documents and throws all tokens into one bag. It is thus natural to take De Finetti's Theorem as inspiration, and as a causal explanation for the existence of the *topics*. Blei et al. (2003) offer a discussion on this. The implication is that in the corpus of news articles on economic expectations, there need to exist latent structures that naturally cluster what is being said and written about in such a way that what is written and said conditional on these clusters is somewhat random. For example, we can have articles discussing oil prices, international politics and healthcare as three different categories – or topics. A measure of evolution in the distribution of these categories, especially in terms of their dispersion over news articles, could then give us a meaningful proxy of the broadness of dialogue – broadness, or a variety of economic beliefs being held at each point in time.

To make the logic of the LDA model clearer, I contrast its likelihood function with other probabilistic generative models in the spirit of Blei et al. (2003). Kindly observe the plate notations[56] in Figure 21.

[55] Similar documents consist of similar terms, and similar terms will be present in similar documents. The most similar columns and rows are compressed into a low-rank representation (k in Figure 4) in the process of matrix factorisation. The factorisation has the power to 'learn' synonymy and polysemy in the language. And this not only in terms of individual words but also entire documents. The latent structure of documents and terms represents almost exactly this – clusters of synonymy, or some degree of linguistic similarity. Aggarwal (2018, p. 38-39) discusses these properties of SVD.

[56] Plate notations are often used in Bayesian machine learning to highlight the workings of the model. The grey nodes represent

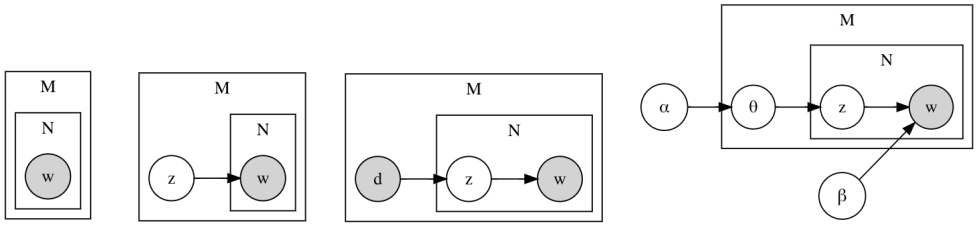


Figure 21: From Left to Right: Unigram Model, Mixture of Unigrams, pLSA and LDA. Source: Own Diagrams.

The corresponding likelihood functions that are assumed to describe the generation of the tokens in a corpus are:

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n) \quad : \text{Unigram Model}$$

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{n} p(w_n|z) \quad : \text{Unigram Mixture Model}$$

$$p(w_n, d) = p(d) \sum_{z} p(w_n|z)p(z|d) \quad : \text{pLSA}$$

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta) \right) d\theta \quad : \text{LDA}$$

where $p(\mathbf{w})$ is the probability of a document, the vector (boldness) of $\mathbf{w}$ represents the fact that a document is a collection of tokens, n is used to denote tokens and z are the topics. The d in pLSA stands for a document, α, β are hyperparameters of Dirichlet distribution and θ is a draw from the Dirichlet. There are several crucial insights to be seen from these equations. In the simplest unigram model, no hidden, latent structure is recognised, since all words are assumed to come from a single multinomial probability distribution. With the mixture model, one assumes there to be a z (topic = mixture component) conditional on which the words are drawn i.i.d. from the respective multinomial distributions. As the z can be seen on the document plate (and not on the word plate), this means that each document is only assumed to have one topic – it is not possible in this simplest mixture model to represent documents as a mixture of topics. When Hofmann (2013) invented pLSA, this was precisely the shortcoming he wanted to alleviate. The pLSA model allows different words coming from different topics, but with the important note that words are i.i.d. only with respect to the *joint* distribution of z, d. The mixture is thus not one defined by a topic, but by the individual combination of topic and document, which means that inference on out-of-sample documents is not obvious, it is prone to over-fitting, and that the number of parameters to be estimated grows *linearly* with the number of documents – which makes the algorithm unpractical for large corpora. The LDA was the first that allowed mixture modelling on both the document and the word level with comparably little computational effort. It was used to construct the Narrative Consensus Index.

Appendix E Intuition: Word Embeddings

An example often made in the literature is the operation of $\mathbf{w}_{\text{king}} - \mathbf{w}_{\text{man}} + \mathbf{w}_{\text{woman}} \approx \mathbf{w}_{\text{queen}}$. The direction and location of the vector for queen will be similar to the vector resulting out of this algebraic operation. This operation is equivalent to thinking in analogies – king is to man what queen is to a woman (or vice versa, a man is to king, what a woman is to queen). See for example Figure 39 in Appendix P, Allen and Hospedales (2019),

observed data; the white nodes are the latent variables to be inferred. Each plate (rectangle) corresponds to a set of data on one hierarchical level, for example, all documents (M, larger rectangle) and all tokens in a document (N, smaller rectangle – subset). Arrows point into the direction of iterative sampling (model generation). Whenever there are nodes outside of plates, one considers these the hyperparameters of the model which are set by the researcher.

the original GloVe paper Pennington et al. (2014), or various internet sources[57] for numerous examples of these analogous relationships discovered by the vector spaces.

The existence of such semantic linear substructures will be used to construct two vectors capturing the expansionary and contractionary narrative. In the following discussion, bear the generic visual image of vector summation and subtraction in mind.

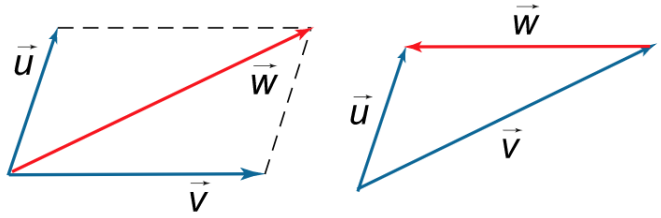


Figure 22: Sum: $\mathbf{u}+\mathbf{v}$ (left) and subtraction: $\mathbf{u}-\mathbf{v}$ (right) of vectors in two-dimensional space. Taken from Svirin (2019).

Let us start with word embedding summation. It has been shown that $\mathbf{w}_{\text{man}}+\mathbf{w}_{\text{royal}} \approx \mathbf{w}_{\text{king}}$ (e.g., Ethayarajh et al., 2019). How can this be? Gittens et al. (2017) very cleverly call this relationship a *paraphrase*. The intuition is that any set of words, let us say a set of context words C where, for example, $C = \{man, royal\}$ are taken, a perfect paraphrase would be a token C_x such that

$$P(W|C) = P(W|C_{\text{x}})$$

where W can be any word in the full vocabulary $\mathcal{V}\backslash\{C, C_x\}$. The intuitive reasoning is that if the meaning of each token is given by a probability distribution over words that co-occur in its vicinity, a perfect paraphrase to a set of words would be a word that has the most similar probability distribution over the contexts as the set of words to be paraphrased. Equivalently, it can be said that when searching for a paraphrase, we are searching for a C_x conditionally on which the observed token co-occurrence probabilities align well with that of $P(W|C_{\text{man}}, C_{\text{royal}})$, $\forall\, w \in \mathcal{V}$ as Gittens et al. (2017) argues. The authors demonstrate that under certain modelling assumptions, which the GloVe model fulfils, and under uniformly distributed marginal word probabilities (unrealistic but needed to simplify the reasoning), the paraphrase of $C_{\text{man}}, C_{\text{royal}}$ is given by the sum of its embeddings. Moreover, since one cannot expect the resulting embedding to exactly match a certain another embedding, a notion of geometrical distance is needed in determining the closest possible paraphrase. Cosine distance (angle) between the paraphrase (summation) vector and other embeddings in the vector space serves to this end. Gittens et al. (2017) show that the cosine distance is indeed a reasonable approach to this since if the angle between any two embeddings $\mathbf{c}_1$ and $\mathbf{c}_2$ is found to be small, of which, e.g., $\mathbf{c}_1$ is the paraphrase embedding, $P(W|c_1)$ and $P(W|c_2)$ will have similar peaks, and thus the cosine distance will convey the similarity in distributions. How is cosine distance and similarity computed?

$$S(\mathbf{u},\mathbf{v}) = 1 - D(\mathbf{u},\mathbf{v}) = cos(\theta) = \frac{\mathbf{u}.\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{17}$$

where S is to denote similarity, D the distance, n is the last element of the vector and θ the angle between the two vectors $\mathbf{u}$ and $\mathbf{v}$. In our case, $\mathbf{u}$ and $\mathbf{v}$ will be the embeddings.

In our concrete example, the addition of the embedding $\mathbf{w}_{\text{royal}}$ aligns the distribution of $P(W|w_{\text{man}})$ to that of $P(W|w_{\text{king}})$ where W represents any and all the words in the vocabulary. The operation *contextualises* word *man*. Remember that geometrically, in terms of its direction (orientation), the sum of two vectors is their average. In the embedding space, this implies an average of two contexts of sorts. In this sense, we have a paraphrase – one can write *king* instead of writing both *man* and *royal* because they have similar context windows. The

[57] Explore for example https://lamyiowce.github.io/word2viz/.

combination of *man* and *royal* literally bears the meaning of a *king*. If we then, for example, sum up vectors for *subprime* and *lending*, we can reasonably expect to be close to *mortgage* since the term subprime lending is overwhelmingly used to refer to a mortgage lending, and thus bears the semantic meaning very similar to *mortgage*. This is, in fact, the case as Figure 39 demonstrates. So by summing up embedded vectors created in vector space models such as GloVe, one is implicitly asking the question: 'Given their context, if we were to observe all of the tokens represented by respective embeddings at once in a text window, what *single* token would be most representative – have the most similar context – of all these *at the same time*?'. Or similarly, 'If $\mathcal{V}$ is the set of all words in the vocabulary of the corpus, and l is the length of the context window from which the term-term co-occurrence matrix is created, what single token w^* at best conveys the meaning of all the tokens in a set $\mathcal{W}$ where $\mathcal{W} \subseteq \mathcal{V}$, $w^* \in \mathcal{V}$ and $|\mathcal{W}| < l$?'. This argumentation is in line with findings of Gittens et al. (2017) and Allen and Hospedales (2019). They call the pair $\{\mathcal{W}, w^*\}$ which forms a paraphrase as *semantically interchangeable* and say that w^* paraphrases $\mathcal{W}$. This logic is shown to generalise to sets of words $\mathcal{W}$ and $\mathcal{W}^*$ where $\mathcal{W}$.

What does it then mean to subtract a vector embedding? Where addition contextualises a word, or narrows its context, vector subtraction will *broaden* the context of a token. Bear in mind that subtraction is just an addition of a vector oriented in the opposite direction. In this sense, a subtraction will discontextualise a word, by removing the context that the subtracted word is described by[58]. This is perhaps at best visualised in the multidimensional space as in Allen and Hospedales (2019) or with a parallelogram, as in Ethayarajh et al. (2019) plotted below.

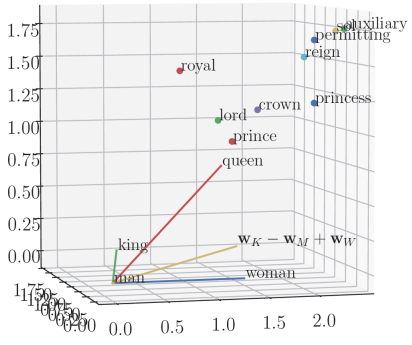


Figure 1. The relative locations of word embeddings for the analogy "*man* is to *king* as *woman* is to ..?". The closest embedding to the linear combination $\mathbf{w}_K - \mathbf{w}_M + \mathbf{w}_W$ is that of *queen*. We explain why this occurs and interpret the difference between them.

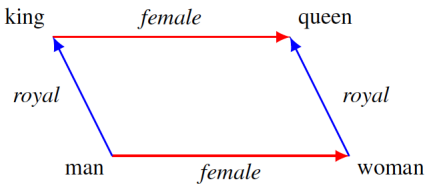


Figure 1: The parallelogram structure of the linear analogy *(king,queen)::(man,woman)*. A linear analogy transforms the first element in an ordered word pair by adding a displacement vector to it. Arrows indicate the directions of the semantic relations.

Analogical Relationships in Vector Space: Taken from Allen and Hospedales (2019) (left) and Ethayarajh et al. (2019) (right)

On the left, one sees the result of the algebraic operation between the three vectors as the yellow vector which is (by cosine distance) closest to $\mathbf{w}_{\text{queen}}$. The locations of the individual words in the space can be thought of as points to which the embeddings of the respective words are exactly pointing. The difference between king and *man* is then the green vector that gets added to the embedding for a woman – the blue line. The embedding of queen is the red line. As you can see, again, there is no reason for why the yellow and red embeddings should be identical; indeed they will be distanced considerably to one another, depending on several terms as identified by Allen and Hospedales (2019). See Appendix F for a discussion of this.

Intuitively, with vector subtraction, the *Euclidian* distance will play an important role. It is important that both pairs, {*king*, *queen*} and {*man*, *woman*} share approximately the same *Euclidian* distance between the words

[58] Alternatively, one can also understand vector embedding subtraction as in Allen et al. (2019), by thinking about a difference in distributions. They show that a vector subtraction results in an embedding that is representative of the Kullback-Leibler divergence between the co-occurrence distributions of the two words. Difference between embeddings is therefore, a measure of meaningful semantic change between two words. If this change can be assumed to be approximately similar, such as here the difference/change in the 'degree' of royalty between man and king, the analogical pair woman and queen is easy to be found because it bears the same semantic difference.

paired in the brackets. If there exists such a semantic 'gender' dimension along which man and woman occur, then the *Euclidian* distance *and* direction between the location of *woman* and *man* are meaningful. Queen is more female than king by the same marginal distance than woman is more female than a man. The direction *and* magnitude of woman − man than approximately gives the gender shift. It can be shown, as Ethayarajh et al. (2019) did, that any analogical relationship approximately satisfies, as Pennington et al. (2014) have conjectured in their original paper, the following

$$\mathbf{w}_{word=a} - \mathbf{w}_{word=b} = \frac{P(W|a)}{P(W|b)} \approx \frac{P(W|x)}{P(W|y)} = \mathbf{w}_{word=x} - \mathbf{w}_{word=y}$$

same as the examined $\mathbf{w}_{king} - \mathbf{w}_{queen} \approx \mathbf{w}_{man} - \mathbf{w}_{woman}$. The proof of this equivalence and approximation rests its argumentation on another proof, provided by Levy, Goldberg, and Dagan (2015), that the GloVe algorithm – same as Word2Vec or older pointwise mutual information (PMI) based approaches to vector space modelling – implicitly factorise a PMI matrix when estimating the embeddings so that

$$\mathbf{w}_i \tilde{\mathbf{w}}_k \;=\; \langle \mathbf{w}_i, \tilde{\mathbf{w}}_k \rangle \;\approx\; \text{PMI}(i,k) = log\left(\frac{P(i,k)}{P(i)P(k)}\right)$$

where i, k are words from $\mathcal{V}$. Factorising a PMI matrix has been shown by Arora et al. (2016) to lead to meaningful, semantic linear substructures in the estimated vector space. It makes sense to realise, at this point, that any analogy can be rewritten as equivalence in paraphrases, that is with summation only, so that the theory above directly applies. In this sense, one can write that if we have an analogy of the type man:king::woman:queen, or a man is to king what a woman is to queen, we can equivalently write any of these three equations

$$\begin{aligned} \mathbf{w}_{\text{king}} - \mathbf{w}_{\text{man}} + \mathbf{w}_{\text{woman}} &\approx \mathbf{w}_{\text{queen}} \\ \mathbf{w}_{\text{king}} - \mathbf{w}_{\text{man}} &\approx \mathbf{w}_{\text{queen}} - \mathbf{w}_{\text{woman}} \\ \mathbf{w}_{\text{king}} + \mathbf{w}_{\text{woman}} &\approx \mathbf{w}_{\text{queen}} + \mathbf{w}_{\text{man}} \end{aligned}$$

One can therefore say that *man transforms* to *king* as *woman transforms* to *queen* from the middle equation, or that {man, queen} paraphrases {woman, king} from the last equation. The interested reader is referred to Ethayarajh et al. (2019) who elaborate on this relationship.

Appendix F Algebraic Operations on Word Embeddings: Approximation Errors

The arguments of Levy et al. (2015), Arora et al. (2016) and Gittens et al. (2017) jointly allow Allen and Hospedales (2019) to explicitly approximate the error of the algebraic operations – explain the difference between the yellow (analogy) and red (queen) embedding in Figure 8 in terms of PMIs. They begin with the definition of paraphrasing. It can then be shown that the following equality holds for any two word sets $\mathcal{W}$ and $\mathcal{W}^*$ where $|\mathcal{W}|, |\mathcal{W}^*| < l$ and most importantly, where $\mathcal{W}$ paraphrases $\mathcal{W}^*$:

$$\sum_{w_i \in \mathcal{W}^*} \mathbf{PMI}_i = \sum_{w_i \in \mathcal{W}} \mathbf{PMI}_i + \boldsymbol{\rho}^{\mathcal{W},\mathcal{W}^*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}^*} - \left(\tau^{\mathcal{W}} - \tau^{\mathcal{W}^*}\right)\mathbf{1}$$

$$\mathbf{w}_{\mathcal{W}^*} = \mathbf{w}_{\mathcal{W}} + \boldsymbol{C}^{\dagger}\left(\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}^*} + \boldsymbol{\sigma}^{\mathcal{W}} - \boldsymbol{\sigma}^{\mathcal{W}^*} - \left(\tau^{\mathcal{W}} - \tau^{\mathcal{W}^*}\right)\mathbf{1}\right)$$

where $\mathbf{w}_{\mathcal{W}} = \sum_{w_i \in \mathcal{W}} \mathbf{w}_i$ and equivalently for $\mathbf{w}_{\mathcal{W}^*}$ is the sum of embeddings. $\boldsymbol{C}^{\dagger}$ is the conjugate transpose of the matrix of the context embedding vectors and $\mathbf{1}$ is a vector of ones. Furthermore, the three error terms ρ, σ and τ are referred to as paraphrase error, conditional independence error and independence error. Elementwise, they can be written as

$$\boldsymbol{\rho}_j^{\mathcal{W},\mathcal{W}^*} = log\frac{P\left(c_j|\mathcal{W}^*\right)}{P\left(c_j|\mathcal{W}\right)} \; c_j \in \mathcal{V}$$

$$\boldsymbol{\sigma}_j^{\mathcal{W}} = log\frac{P\left(\mathcal{W}|c_j\right)}{\prod_i P\left(w_i|c_j\right)}$$

$$\tau^{\mathcal{W}} = log\frac{P(\mathcal{W})}{\prod_i P(w_i)}$$

Furthermore, Allen and Hospedales (2019) argue that if words (tokens) have similar dependence terms (are in substantial way semantically similar), the terms $\boldsymbol{\sigma}^{\mathcal{W}}, \boldsymbol{\sigma}^{\mathcal{W}^*}$ and $\tau^{\mathcal{W}}, \tau^{\mathcal{W}^*}$ tend to cancel out (or their difference be relatively small). In that situation, the paraphrase error $\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}^*}$ is then the sole substantial source of error. Regrettably, it is neither mathematically straightforward nor feasible in our model to obtain an estimate of $\boldsymbol{\rho}^{\mathcal{W},\mathcal{W}^*}$. Nonetheless, it seems reasonable to assume that even if this error should be considerable in size, if the algebraic vector operation is defined sensibly, the resulting word embedding will point to the desirable semantic direction nevertheless. As Arora et al. (2016) points out, the low dimensionality of the vector embeddings has a 'purifying' effect, reducing the error associated with the error terms here. It is this reasoning which underlies much of the motivation for using a dimensionality-reducing algorithm such as GloVe to find the low-rank approximation of the term-term-co-occurrence matrix, instead of using the original matrix. The need to compress the meaning of a word to 100 numbers, instead of hundred-thousands (as in the original count-based co-occurrence matrix $\mathbf{X}$) has the effect of identifying patterns and relationships between words that could otherwise not been identified.

Appendix G Narrative Lexicons

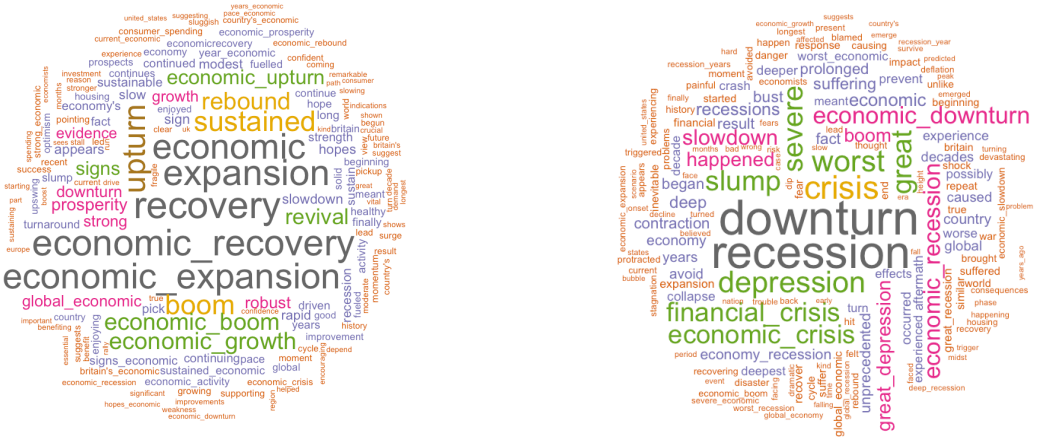


Figure 23: Expansion (left) and Contraction (right) Paraphrase: Closest Words

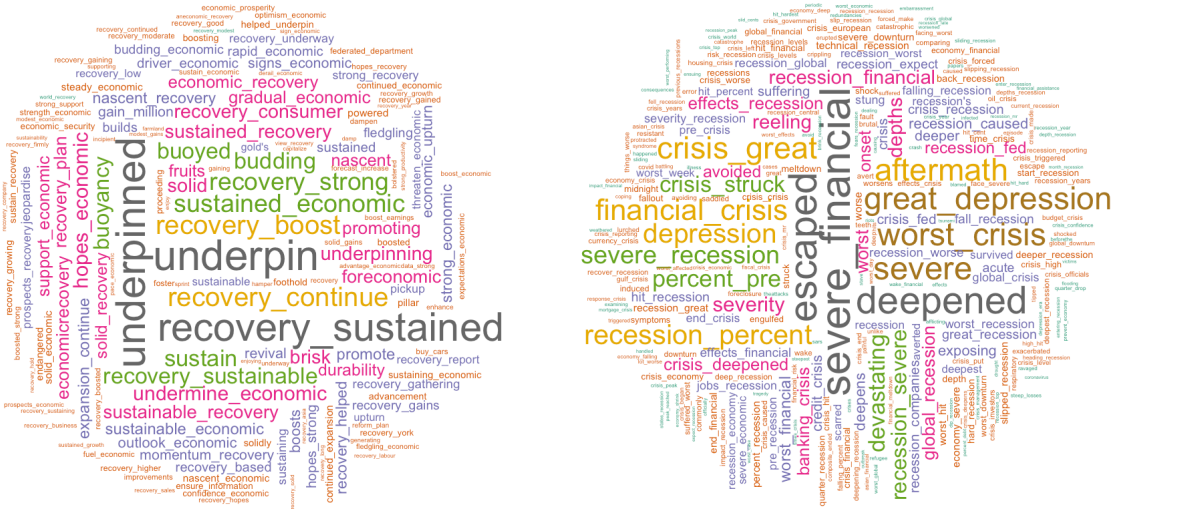


Figure 24: Difference: Expansion - Contraction (left) and Contraction - Expansion (right) : Closest Words

Table 13: The Expansionary (Narrative) Lexicon

Ranking	Word	Cosine Distance	Ranking	Word	Cosine Distance	Ranking	Word	Cosine Distance
1	underpinned	0.4896419	71	optimism_economic	0.3457394	141	pace_economic	0.3093429
2	solid	0.4729543	72	enjoy	0.3449835	142	improvement	0.3093144
3	underpin	0.4523427	73	economic_prosperity	0.3447595	143	budding_economic	0.3092853
4	economic_recovery	0.439903	74	dim	0.3434108	144	showing_strong	0.3082014
5	recovery_sustained	0.4370506	75	recovery_based	0.3424596	145	recovery_broad	0.3080669
6	sustained	0.4329551	76	expectations_economic	0.3416067	146	recovery_gains	0.3078852
7	buoyed	0.4329541	77	enjoying	0.3398207	147	torise	0.3078702
8	sustained_economic	0.4319943	78	fundamentals	0.3396912	148	recovery_good	0.3068163
9	outlook_economic	0.4297857	79	sustainability	0.3386506	149	recovery_higher	0.3067888
10	strong_economic	0.4230501	80	strengthening	0.3381375	150	reinforce	0.3064066
11	recovery_continue	0.4219974	81	strong_buy	0.3369635	151	gradual	0.3063361
12	underpinning	0.4185569	82	confidence_economic	0.3366517	152	recovery_boosted	0.3063321
13	recovery_strong	0.4143061	83	driver_economic	0.3363872	153	data_strong	0.3053107
14	upturn	0.4134577	84	prosperity	0.3362308	154	pillar	0.3048931
15	economicrecovery	0.4101162	85	momentum	0.3354046	155	enjoyed	0.3039843
16	sustain	0.4096327	86	strong_growth	0.3349149	156	continue	0.3035775
17	economic_upturn	0.4057397	87	lifted	0.3342415	157	sustain_economic	0.3034309
18	boosted	0.4056508	88	sustained_growth	0.3330094	158	evidence_economic	0.3034097
19	strong	0.394067	89	sustainable_recovery	0.3329834	159	view_recovery	0.3020425
20	hopes_economic	0.393274	90	maintaining	0.3324509	160	foster	0.3019185
21	support_economic	0.3928549	91	hopes_strong	0.3324166	161	buy_cars	0.3017378
22	recovery_consumer	0.3885712	92	moderate	0.332189	162	recovery_year	0.3016855
23	revival	0.3882171	93	strength	0.3318159	163	based_investment	0.3012839
24	promoting	0.3877773	94	economic_growth	0.3314699	164	supports	0.300452
25	expansion_continue	0.3857539	95	continued_expansion	0.3312542	165	firmly	0.3001247
26	recovery_boost	0.3856984	96	benefiting	0.331045			
27	boosting	0.3856115	97	gaining	0.3308948			
28	recovery	0.3847741	98	recovery_helped	0.3303784			
29	signs_economic	0.3838074	99	german_economic	0.3300657			
30	sustainable	0.3834572	100	fruits	0.3296682			
31	promote	0.3791704	101	fledgling	0.3268559			
32	pickup	0.379161	102	generating	0.326811			
33	sustained_recovery	0.37771	103	solid_growth	0.3263873			
34	boosts	0.3754686	104	boosted_strong	0.3261277			
35	buoyancy	0.3748099	105	endangered	0.3260251			
36	supporting	0.3716398	106	secure	0.3256979			
37	solidly	0.3689898	107	expansion	0.323361			
38	budding	0.3664895	108	nascent_recovery	0.3230609			
39	solid_economic	0.3658884	109	rapid	0.3224757			
40	rapid_economic	0.3653676	110	underway	0.3219747			
41	brisk	0.3634908	111	recovery_york	0.3203561			
42	momentum_recovery	0.3633626	112	sign_economic	0.3202617			
43	continued_economic	0.3626961	113	hamper	0.3187747			
44	prospects_recovery	0.3619977	114	healthy	0.3182447			
45	prospects	0.3613961	115	gold's	0.3182329			
46	improvements	0.3611093	116	foothold	0.3180793			
47	robust	0.3610718	117	recovery_firmly	0.3177528			
48	dampen	0.3601053	118	encouraged	0.3176399			
49	undermine_economic	0.359733	119	improving	0.3175961			
50	foreconomic	0.3582221	120	solid_gains	0.3162679			
51	strong_recovery	0.3581229	121	subdued	0.3162247			
52	modest	0.3580447	122	strong_support	0.316218			
53	steady	0.3568774	123	encouraging	0.3156296			
54	bolstered	0.3553822	124	recovery_low	0.3156246			
55	builds	0.3542517	125	boost_economic	0.3156149			
56	nascent	0.354118	126	hopes	0.3149665			
57	prospects_economic	0.35407	127	nascent_economic	0.3143098			
58	stable	0.3537585	128	enhance	0.3131611			
59	recovery_sustainable	0.3527833	129	threaten_economic	0.3130646			
60	sustaining	0.3526994	130	essential	0.3127519			
61	solid_recovery	0.3522645	131	supported	0.3122421			
62	gradual_economic	0.3517809	132	gains	0.3122215			
63	strength_economic	0.3504985	133	benefited	0.3113127			
64	durability	0.3504946	134	supportive	0.3106368			
65	economic_expansion	0.3503754	135	recovery_modest	0.3104049			
66	recovery_underway	0.3500299	136	modest_economic	0.3102706			
67	sustainable_economic	0.3476242	137	recovery_growth	0.3101217			
68	jeopardise	0.3472059	138	gain_million	0.3101142			
69	steady_economic	0.3470591	139	recovery_gaining	0.3100048			
70	recovery_plan	0.3460113	140	ensure_information	0.3097458			

Table 14: The Contractionary (Narrative) Lexicon

Ranking	Word	Cosine Distance	Ranking	Word	Cosine Distance	Ranking	Word	Cosine Distance
1	severe	0.5468059	92	meltdown	0.3540041	183	recession_years	0.3174437
2	severe_financial	0.5239225	93	severity_recession	0.3537833	184	respiratory	0.317273
3	escaped	0.5159662	94	stung	0.3535949	185	coronavirus	0.3167338
4	deepened	0.502152	95	scarred	0.3530837	186	dip	0.316729
5	financial_crisis	0.4990616	96	crises	0.3527248	187	written	0.3167213
6	worst	0.4980525	97	avoiding	0.3522901	188	negative_effect	0.3164516
7	depression	0.4939162	98	cases	0.3497142	189	technically	0.3156322
8	great_depression	0.4880091	99	deepest	0.3489245	190	economy_recession	0.3155504
9	worse	0.4649676	100	effects_crisis	0.3477235	191	responded	0.3155429
10	aftermath	0.4644937	101	recession_expect	0.3472256	192	causing	0.3154336
11	recession	0.4639323	102	exposing	0.3469634	193	infected	0.3153838
12	devastating	0.4627149	103	negative_territory	0.3465225	194	financial	0.3152097
13	crisis_struck	0.4581345	104	severe_downturn	0.3455216	195	outbreak	0.3151752
14	severe_recession	0.4533966	105	commonly	0.342604	196	deflation	0.3151562
15	crisis_great	0.4483207	106	hit_percent	0.3421594	197	crisis_put	0.3149837
16	crisis	0.4483141	107	territory	0.3420406	198	sars	0.3142867
17	worst_crisis	0.4445437	108	dealing	0.3405394	199	crisis_levels	0.3141005
18	suffering	0.44228	109	exacerbated	0.3396701	200	recession_quarter	0.3140559
19	reeling	0.4415779	110	crisis_economy	0.3391504	201	harsh	0.3139301
20	deeper	0.4362103	111	resistant	0.3382706	202	economic_crisis	0.3137625
21	recession_severe	0.4357483	112	economy_financial	0.33779	203	teeth	0.3135267
22	avoided	0.4352971	113	global_financial	0.337692	204	shortly	0.3132398
23	effects_recession	0.4312559	114	adverse	0.3373846	205	crisis_investors	0.3131724
24	recession's	0.4264553	115	start_recession	0.3372926	206	felt	0.3130999
25	effects	0.4262268	116	time_crisis	0.3372048	207	hit_hard	0.3130437
26	severity	0.4238936	117	deep_recession	0.3369753	208	drastic	0.3125885
27	recession_percent	0.4216328	118	crisis_worse	0.3366193	209	impact	0.3124945
28	shock	0.4213915	119	worst_affected	0.3353388	210	losses	0.3123082
29	technical_recession	0.4154653	120	end_financial	0.3352283	211	repeat	0.3119778
30	negative	0.4102299	121	symptoms	0.3350761	212	crippling	0.3117982
31	struck	0.4092103	122	brutal	0.3345507	213	tipped	0.3117618
32	depths	0.407588	123	crisis_hit	0.3344869	214	economy_crisis	0.3117133
33	global_recession	0.4064802	124	waves	0.3336521	215	suffer	0.3114776
34	downturn	0.406091	125	feared	0.3331268	216	blamed	0.3111897
35	happened	0.4041917	126	bad	0.3319379	217	currency_crisis	0.3110637
36	effects_financial	0.40414	127	occurred	0.3319303	218	risk_recession	0.3097575
37	percent_pre	0.4017754	128	falling_percent	0.3316047	219	plunged	0.3097238
38	great	0.4008615	129	escape	0.3311224	220	face_severe	0.3097112
39	recession_financial	0.400767	130	hits	0.3310004	221	disaster	0.3096307
40	back_recession	0.3970706	131	recession_global	0.3306415	222	worst_economic	0.3095243
41	recessions	0.3955872	132	recession_great	0.3306021	223	nasty	0.3091291
42	great_recession	0.3938162	133	crisis_financial	0.3304976	224	unknown	0.3090521
43	acute	0.3917572	134	recession_reporting	0.3302154	225	negative_outlook	0.3089372
44	recession_companies	0.3902781	135	facing_worst	0.3297093	226	slip_recession	0.3087754
45	depth	0.3890411	136	recession_levels	0.3294489	227	induced	0.3080097
46	suffered	0.38785	137	hard_recession	0.3293156	228	turned	0.3076438
47	onset	0.3870092	138	crisis_european	0.3292976	229	hit	0.3075697
48	negative_effects	0.3860166	139	avert	0.329275	230	disastrous	0.3074985
49	dire	0.3858678	140	global_crisis	0.3290444	231	financial_risk	0.3074254
50	worst_week	0.3843597	141	depth_recession	0.3280417	232	technical	0.3068852
51	affected	0.3839175	142	asian_crisis	0.3278947	233	matters	0.3068237
52	consequences	0.3832481	143	lot_worse	0.32738	234	economic_downturn	0.306524
53	fallout	0.3816575	144	negative_growth	0.3268927	235	trigger	0.3064949
54	severe_economic	0.3812571	145	worst_downturn	0.3266219	236	situation_worse	0.3063382
55	worst_financial	0.379938	146	dip_recession	0.3265641	237	crisis_crisis	0.3061335
56	banking_crisis	0.3798047	147	fell_recession	0.3265548	238	percent_recession	0.3059071
57	recession_fed	0.3797634	148	worst_day	0.3262984	239	inevitable	0.30577
58	impact_recession	0.3794188	149	contraction	0.3261835	240	previous	0.3057211
59	things_worse	0.3776501	150	error	0.3261596	241	crisis_level	0.3052239
60	economy_severe	0.3773493	151	crisis_recession	0.3255025	242	depths_recession	0.3050652
61	recession_economy	0.3772971	152	deepening_recession	0.3252372	243	protracted	0.3049251
62	recession_worst	0.3771329	153	bracing	0.3249996	244	turns	0.3047486
63	catastrophic	0.3759333	154	papers	0.3248518	245	economy_deep	0.304665
64	credit_crisis	0.3748513	155	triggered	0.324683	246	contagion	0.3045492
65	crisis_deepened	0.3736503	156	catastrophe	0.3244216	247	crisis_made	0.3045289
66	worst_recession	0.3723826	157	recession_year	0.3243656	248	crisis_high	0.3045183
67	wake	0.3716277	158	immediately	0.324035	249	response	0.3042507
68	suffered_worst	0.3704664	159	sliding	0.3238015	250	heading_recession	0.3041509
69	hit_recession	0.3689601	160	panic	0.3236425	251	immune	0.3036593
70	pre_crisis	0.3684945	161	end_crisis	0.3233962	252	fault	0.3034062
71	fall_recession	0.3682873	162	quarter_recession	0.3230423	253	deepest_recession	0.303164
72	comparing	0.3680766	163	previous_recessions	0.3229317	254	crunch	0.3031143
73	avoid	0.3676506	164	similar	0.3225475	255	covid	0.3029205
74	falling_recession	0.3670725	165	virus	0.3225437	256	fared	0.3026859
75	worst_hit	0.3668926	166	worst_effects	0.322447	257	feedback	0.3026734
76	pre_recession	0.3649089	167	crash	0.3219312	258	turn_worse	0.3019855
77	recession_worse	0.3626353	168	worst_performing	0.321896	259	recovered	0.3019478
78	unlike	0.3621724	169	recession_central	0.3214404	260	sliding_recession	0.3018722
79	deeper_recession	0.3619659	170	saddled	0.3213378	261	terrible	0.3017085
80	worsened	0.3613375	171	economy_falling	0.3212777	262	housing_crisis	0.3016591
81	survived	0.359411	172	worst_case	0.3210856	263	severely	0.3015414
82	deepens	0.3584777	173	shocked	0.3210287	264	engulfed	0.3014559
83	slipped_recession	0.3571363	174	midnight	0.3207472	265	hit_hardest	0.3013427
84	jobs_recession	0.3571276	175	completely	0.3200651	266	cumulative	0.3008717
85	averted	0.3569458	176	comparison	0.3197897	267	crisis_officials	0.3005347
86	caught	0.3568022	177	worsens	0.3197257	268	officially	0.3003288
87	caused	0.3562832	178	slipping_recession	0.3195822	269	impact_financial	0.3003283
88	deep	0.3561921	179	turn_negative	0.3186423	270	debt_crisis	0.3001794
89	crisis_fed	0.3561629	180	plausible	0.3179225	271	episode	0.3001042
90	recession_caused	0.3561165	181	percent_worst	0.3177664	272	slump	0.3000337
91	painful	0.3549204	182	recover_recession	0.3177215			

Appendix H Time Series Modelling and Statistical Testing

H.1 Scaling, Filtering and Smoothing

When plotting data, to perform visual evaluation and to provide analytical insights, standard scaling, filtering, and smoothing shall be applied. To ensure graphic comparability of two time series with substantially different ranges, the time series are normalised with their standard deviation, so that they have $\mu = 0$ and $\sigma^2 = 1$ when presented in comparison charts. Mathematically, the z-score of each observation is computed:

$$x_i^* = \frac{x_i - \mu}{\sigma}$$

where x_i^* is the scaled series. The direction and variability in the relative movements of the scaled series are so readily visually comparable, even between differently-scaled time series.

In order to strengthen analytic insights, and especially to highlight trends, when plotting the text-based indices, I will enrich them with HP-filtered values. Since either because of size limitations of the corpus or because of marked tendencies to change wording abruptly, the series varies substantially in monthly intervals; the resulting indices would perhaps appear too noisy without a filter. Moreover, the trend of the indices could harbour hints about the just transpiring business cycle stage, and filtering highlights such a trend. Hodrick-Prescott filter shall be used. It decomposes a time series into a trend and a cyclical component by minimising the loss function defined with

$$\min_{\tau} \left(\sum_{t=1}^{T} (y_t - \tau_t)^2 + \sum_{t=1}^{T} \lambda \left(\tau_{t+1} - \tau_t\right) \left(\tau_t - \tau_{t-1}\right) \right)$$

where τ_t is the trend component y_t is the observed time series to be filtered and $y_t = c_t + \tau_t + \epsilon_t$ with c_t being the cyclical component and ϵ_t an estimation residual. The first term penalises the deviation of the trend component from the observed time series and the second penalises the deviations of the trend component from a stable growth path. $\lambda = 800$ was used. For more information, cf. Hodrick and Prescott (1997).

Lastly, double exponential smoothing of the indices was plotted to look for interesting patterns. The procedure is known as Holt-Winters double exponential smoothing based on Holt (2004) and Winters (1960) was used. The method creates the smoothed values as follows

$$\begin{aligned} s_1 &= y_1 \\ b_1 &= y_2 - y_1 \\ s_t &= \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} \end{aligned}$$

where $\{s_t\}$ is the smoothed time series, $\{b_t\}$ is the best estimate of a trend at time t, α is the data smoothing factor and is together with β, the trend smoothing factor, exogenously given. $\alpha = 0.3$ and $\beta = 0.5$ were used so that the series reacts to new values fairly quickly and markedly. The double exponential smoothing procedure is to be differentiated from simple single exponential smoothing by its explicit introduction of a trend in calculating the smoothed series – $\{b_t\}$ is basically a smoothed moving trend. Importantly, note that contrary to the HP-filter, the double smoothing procedure does not include future values in the smoothed value. The filters are used for purely descriptive, visual purposes.

H.2 Correlations and Correlation Functions

We are interested in correlation coefficients between the text-based indices and the comparable time series. This not only in equivalent periods, but with relative differences in periods (lags and leads) of the comparative series. For this purpose, cross-correlation functions are handy. Correlation and correlation function respectively is defined as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$g_k^{xy} = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})\,(x_{t+k} - \bar{x})$$

where x_i and y_i are the observations i of the time series X and Y, with their respective means denoted by bars. The first function calculates the regression coefficient, the latter the correlation function for a given lag or lead k. If we are to expect predictive power of the text-based indices towards the comparative time series, y being the indices and x the comparative series, then the correlations should be significant where $k > 0$ – on the lead side of the correlation function, when the function pinpoints correlation between y_t and x_{t+k}.

H.3 Regressions and In- and Out-of-Sample Forecasts

In order to assess the predictive value of the text-based indices, a wide array of possible regression models was examined. Among others, simple regressions of the forms below were fitted:

$$GDP_t = \alpha + \beta_1 RS_{t-1} + \epsilon_t$$
$$GDP_t = \alpha + \beta_1 GDP_{t-1} + \beta_2 RS_{t-1} + \epsilon_t$$
$$GDP_t = \alpha + \beta_1 RS_{t-1} + \beta_2 RS_{t-2} + ... + \beta_p RS_{t-p} + \epsilon_t$$
$$GDP_t = \alpha + \beta_2 RS_t + ... + \beta_p RS_{t-p} + \gamma_1 Entropy_t + ... + \gamma_r Entropy_{t-r} + \epsilon_t$$
$$GDP_t = \alpha + \beta_1 RS_{t-1} + \beta_2 RS_{t-2} + ... + \beta_p RS_{t-p} + \gamma_1 Entropy_{t-1} + ... + \gamma_r Entropy_{t-r} + \epsilon_t$$
$$GDP_t = \alpha + \beta_2 RS_{t-2} + ... + \beta_p RS_{t-p} + \gamma_1 Entropy_{t-2} + ... + \gamma_r Entropy_{t-r} + \epsilon_t$$

where GDP_t stands for nominal, seasonally adjusted, GDP QoQ growth. The last three lines are used to construct current period, one-period ahead, and two-period ahead forecasts of nominal GDP QoQ growth in the U.S. These equations mostly resemble the group of time series models referred to as Autoregressive Distributed Lag (ADL) models. Significance of the coefficients, R squared value, and the F-statistic of the model's coefficients are examined. It is clear that these regressions potentially suffer from substantial omitted variable bias. This issue should, however, not discourage us, since proving causality is not the aim. The primary purpose of this exercise is to show the predictive power of the indices. Afterwards, with the estimated model parameters, in-the-sample and out-of-sample forecasts are created. When out-of-sample forecasts are created, a cut-off is defined to create a subset of the time series to train the model on, and subsequently predict the pseudo-future values. Prediction success – the correctness of the sign, trend and level – is plotted in tables. The predictions are also evaluated with a measure known as root mean square error. RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{t=t_0}^{T}(\hat{y}_t - y_t)}{T - t_0 + 1}}$$

where I used t_0 to denote some starting period at which the first in- or out-of-sample forecast was made, and T denotes the end period. $\hat{y}_t$ stands for the forecasted value and y_t is the actually observed value.

H.4 Granger Causality

Related to the standard VAR analysis found in much of macroeconomic literature, is the concept and the statistical testing of Granger causality. Granger (1969) asked the question whether a time series can help to forecast another one, and introduced a specific bivariate framework to test that hypothesis which much resembles the standard VAR. Let us take, for example, a bivariate regression of GDP_t on its lags and the lags of RS_t and vice versa. GDP_t is defined as in Section 4.5. Including intercepts, the equations would take the following form:

$$GDP_t = c_1 + \alpha_1 GDP_{t-1} + \alpha_2 GDP_{t-2} + ... + \alpha_p GDP_{t-p} + \beta_1 RS_{t-1} + \beta_2 RS_{t-2} + ... + \beta_p RS_{t-p} + \epsilon_t$$
$$RS_t = c_2 + \gamma_1 GDP_{t-1} + \gamma_2 GDP_{t-2} + ... + \gamma_p GDP_{t-p} + \delta_1 RS_{t-1} + \delta_2 RS_{t-2} + ... + \delta_p RS_{t-p} + \epsilon_t$$

After estimating the first equation – coefficients α and β – with OLS, a Wald test with the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ can be performed. The same is done for the second equation. The Wald statistic can then be calculated by obtaining the residual sum of squares from the bivariate (unrestricted) model and the corresponding residual sum of squares from the univariate (restricted) model where H_0 applies. The corresponding statistic is defined as:

$$F = \frac{(RSS_u - RSS_r)/p}{RSS_r/(T - 2p - 1)}$$

where RSS_u and RSS_r are the residual sums of squares of the unrestricted and restricted model respectively, and $T - 2p - 1$ are the degrees of freedom. If the F-statistic turns out significant in the first model specification, we could say that RS_t Granger-causes GDP_t. If it should be the case in the second specification, we could say that GDP_t Granger-causes RS_t. The ideal case to lend support for the predictive value of the indices would be to test significant on the former, and insignificant on the latter. Notice that Granger causality in both directions could be the case. It should be stressed at this point, that even though the name of the test bears 'causality', the result in no way proves a causal relation. At best, it makes it more probable. The exposition here is taken from J. D. Hamilton (1994).

Much of Granger's work and the statistical tests surrounding it, however, prerequisite the time series at hand to be neither cointegrated nor non-stationary (not integrated), so that the associated asymptotic theory holds. This has been shown, i.a., by Sims, Stock, and Watson (1990). As is common in macroeconomics, and even more so in case of seasonally adjusted GDP growth, several of time series analysed here cannot be expected to be stationary. Therefore, the conclusions based on the Granger causality test could be incorrect. Thankfully, Toda and Yamamoto (1995) have developed a framework to circumvent this problem with an approach that allows Granger causality testing adhering to standard asymptotic theories no-matter the order of integration, as long as one possesses the information on the maximum order of integration, takes care of serial autocorrelation of error terms, and the 'optimal' VAR constructed – as chosen by an information criterion – has lag order equal or greater than the maximum order of integration of the time series. The steps to perform this test follow:

1. Estimate the maximum order of integration (d_{max}) by means of unit root tests. In the case no unit-roots are found, the standard Granger causality testing may be performed. Kindly cf. the subsection on robustness checks.
2. Construct a VAR model in levels (no integration) where an information criterion determines the number of lags. Akaike Information Criterion (AIC) was used. Denote the order of lags by p.
3. Check for residual autocorrelation with appropriate tests such as the Breusch-Godfrey test (cf. Appendix I, subsection on serial correlation). If serial autocorrelation is found, add lags until it disappears. This step could result in additional lags $p_{auto} \geq 0$.

4. Add the number of lags that you found the time series to be integrated d_{max}. The total number of lags is now $p + p_{auto} + d_{max}$.
5. Finally, perform a Wald test on the coefficients of the lags of the variables that you hypothesis would Granger-cause the dependent variable in the respective equation. Only perform the test on the p lags (without p_{auto} or d_{max}) with associated degrees of freedom.

The Granger causality testing procedure of Toda and Yamamoto (1995) is performed on the bivariate regression of each of the two constructed time series (RS_t and $Entropy_t$) respectively regressed on each of the comparative time series such as GDP_t, EPU_t, etc. individually. The resulting Wald tests and inferences are communicated in tables. Testing in both directions is performed.

H.5 Structural Break Analysis

Lastly, we want to evaluate when significant trend shifts in the indices occurred, and whether these shifts correspond with actual business cycle turning points. This finding would lend support to the interpretation that shifts in the index could bear macroeconomic importance and be the essence of evaluating early crisis warning potential of the indices. There are many econometric methods for identifying structural breaks, but most of them focus on testing either exogenously given breaks, or only singular breaks. For the purposes here, we will need a method that is able to extract multiple structural breaks.

Zeileis, Kleiber, Krämer, and Hornik (2003) provide insight into structural break testing and implement the R package *strucchange* used for the analysis here. Testing for structural breaks in time series linear regression models can be thought of as testing the null hypothesis of constant regression coefficients such that $H_0 : \beta_i = \beta_0 \ \forall i$ where $i = \{1, ..., n\}$. The alternative hypothesis is then that there exists (at least one) coefficient β, where the null hypothesis does not hold. This method boils down to creating linear regression models on $m \geq 1$ partitions of the domain (time) and testing their relative ability to correctly capture the data by a metric such as the residual sum of squares. Mathematically, this can be expressed as a test on the linear regression models of the form

$$\begin{aligned} y_i &= x_i^T \beta_i + \epsilon_t \ \text{ with } i = \{1, ..., n\} && : H_0 \text{ conform model} \\ y_i &= x_i^T \beta_j + \epsilon_t \ \text{ with } j = \{1, ..., m\} \text{ and } i = \{i_{j-1} + 1, ..., i_j\} && : H_A \text{ conform model} \end{aligned}$$

where j is the set of identified breakpoints where the time series is partitioned. The procedure boils down to a test of an unrestricted vs. restricted setting (F-test or Wald statistic) of the null vs. alternative models. In an early method of finding unknown structural breaks, Andrews (1993) has suggested this F statistic be computed as (based on Zeileis et al. (2003))

$$F_s = \frac{\hat{u}^T \hat{u} - \hat{u}(s)^T \hat{u}(s)}{\frac{\hat{u}(s)^T \hat{u}(s)}{n-2k}}$$

where s will be a specific time period at which the partition occurs so that $s = \{n_h, ..., n - n_h\}$, $n_h \geq k$ and $n_h = \lfloor nh \rfloor$ which will be some minimum number of periods to be included in any resulting segment. Bai and Perron (2003) have contributed to expanding this F-statistic to explicit testing of *multiple* breaks as an alternative hypothesis. Moreover, they developed an algorithm to minimise the residual sum of squares of the resulting partitioned model based on the Bellman principle that makes the finding of multiple breakpoints mathematically exact and computationally feasible. The method of Bai and Perron (2003) is used here. The optimal number of partitions is estimated with BIC. The minimum amount of time periods contained within a segment is based on historical evidence – 6 months ($\approx$ two quarters). The model to be tested will be a simple regression of the

two indices, RS_t and $Entropy_t$ individually on their intercepts. A structural break in this model discovers the most sizable shifts in the mean of the indices. These should coincide or even predict actual business cycle turning points as defined by National Bureau of Economic Research (n.d.). For further relevant econometric tests and context of their results, cf. Appendix I.

Appendix I Further Econometric Tests

Many of the methods discussed in Section 4.5 and Appendix H require additional testing in order for their results and conclusions to be correct. The credibility of the results would also benefit from further cross-checking. This section shortly discusses the type of tests that were performed in order to confirm the validity of the results obtained above.

I.1 Unit Root Tests

Whenever OLS regression was estimated, when interpreting the estimated coefficients, we should make sure the resulting time series are stationary. If that is not the case, cointegration should be examined so that the estimated regression is not spurious. The concept of stationarity is equivalent to refuting the presence of a unit root (a random walk process). There are many tests used to this end in the literature, of which I shall make use of the Augmented-Dickey-Fuller (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KSS) test (cf. Dickey & Fuller, 1979, 1981; Kwiatkowski, Phillips, Schmidt, & Shin, 1992). The former tests the null hypothesis of a unit root with the alternative of a stationary process and the contrary applies to the KSS. We, therefore, seek to reject the null hypothesis of the ADF test and not reject the null hypothesis of KSS. The results are provided below.

Table 15: ADF-Tests: Relative Sentiment and Entropy Index

		Unit Root	Drift and Unit Root
Null Hypothesis: Relative Sentiment (monthly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
Augmented Dickey-Fuller test statistic		-4.1814^{***}	8.8731^{***}
Test critical values:	1% level	-3.44	6.47
	5% level	-2.87	4.61
	10% level	-2.57	3.79
Null Hypothesis: Relative Sentiment (quarterly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		-2.3554	3.069
Test critical values:	1% level	-3.46	6.52
	5% level	-2.88	4.63
	10% level	-2.57	3.81
Null Hypothesis: Entropy (monthly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		-4.2676^{***}	9.1165^{***}
Test critical values:	1% level	-3.44	6.47
	5% level	-2.87	4.61
	10% level	-2.57	3.79
Null Hypothesis: Entropy (quarterly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		-3.1998^{**}	5.1376^{**}
Test critical values:	1% level	-3.46	6.52
	5% level	-2.88	4.63
	10% level	-2.57	3.81

Note: *p<0.1; **p<0.05; ***p<0.01 . Test result is confirmed by a Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Not reported for brevity.

Table 16: ADF-Tests: Comparative Time Series

Null Hypothesis: GDP Growth (quarterly, nominal) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		−1.3162	1.3398
Test critical values:	1% level	-3.46	6.52
	5% level	-2.88	4.63
	10% level	-2.57	3.81
Null Hypothesis: GDP Growth (quarterly, real) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		−1.5101	1.5373
Test critical values:	1% level	-3.46	6.52
	5% level	-2.88	4.63
	10% level	-2.57	3.81
Null Hypothesis: VIX (monthly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		−5.2551***	13.8361***
Test critical values:	1% level	-3.44	6.47
	5% level	-2.87	4.61
	10% level	-2.57	3.79
Null Hypothesis: UMCSENT (monthly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		−3.1099**	4.8619**
Test critical values:	1% level	-3.44	6.47
	5% level	-2.87	4.61
	10% level	-2.57	3.79
Null Hypothesis: EPU (monthly) has a unit root			
Type: Drift (Constant)			
Lag Length: 1 (Automatic - based on BIC)			
		Unit Root	Drift and Unit Root
Augmented Dickey-Fuller test statistic		−4.2414***	9.1382***
Test critical values:	1% level	-3.44	6.47
	5% level	-2.87	4.61
	10% level	-2.57	3.79

Note: *p<0.1; **p<0.05; ***p<0.01. The test result is confirmed by a Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Not reported for brevity.

I.2 Portmonteau and Breusch-Godfrey LM Statistic

When performing Granger causality testing according to Toda and Yamamoto (1995), it was important to make sure that the residual terms from the final bivariate model be not serially autocorrelated. The VAR models estimated with Granger causality testing are evaluated with the Portmanteau statistic (Q-Tests, or Box-Pierce and Ljung-Box) and the Breusch-Godfrey LM-statistic. For further detail, you can consult Breusch (1978), Godfrey (1978) or J. D. Hamilton (1994).

Table 17: Serial Correlation Testing

VAR Model	Lags Included	Portmonteau	DF	P-Value	Breusch-Godfrey LM Test	DF	P-Value
RS & GDP Growth (nom)	1 (0)	28.452	60	0.9999	14.812	20	0.787
RS & GDP Growth (real)	1 (0)	29.333	60	0.9997	14.653	20	0.7959
RS & VIX	7 (0)	33.266	36	0.5993	19.83	20	0.4686
RS & UMCSENT	6 (0)	39.39	40	0.4976	17.62	20	0.6124
RS & EPU	9 (0)	31.961	28	0.2761	18.145	20	0.5779
Entropy & GDP Growth (nom)	1 (0)	31.831	60	0.9998	12.716	20	0.787
Entropy & GDP Growth (real)	1 (0)	33.886	60	0.9974	13.729	20	0.844
Entropy & VIX	3 (0)	35.916	52	0.9564	22.949	20	0.2913
Entropy & UMCSENT	2 (0)	50.534	56	0.6811	27.49	20	0.122
Entropy & EPU	4 (9)	19.158	12	0.08479	15.685	20	0.736

Note: This table reports the results of serial correlation tests according to Breusch (1978) and Godfrey (1978) while performing the necessary testing in the course of Granger causality testing with the approach of Toda and Yamamoto (1995). The null hypothesis is that of *no* serial correlation present. In lags included, the first number stands for the AIC chosen VAR model, and the number in parentheses stands for the number of lags needed to be included until no serial correlation was present. For *Entropy* with *EPU*, there was no specification at which both tests turned out insignificant. *p<0.1; **p<0.05; ***p<0.01

I.3 Cointegration

The concept of cointegration is crucial whenever a series is integrated while performing regression analysis. Ever since the works of Yule (1927) and Granger and Newbold (1974), it is clear that two time series might appear to be correlated even though there is no actual real relationship between them. Also concerning is the fact that standard asymptotic theory needed for inference on OLS coefficients is non-trivial, and thus standard hypothesis testing should not be performed. If two series are spuriously correlated, coefficients estimated with OLS will be inconsistent. If they are cointegrated, it can be argued that there exists a long-term relationship between the variables, as the series follow the same stochastic trend, and the estimated coefficients will be (super)consistent. Even in the presence of cointegration, the inference is still not trivially possible. Approaches such as the Engle-Granger two-step method or Autoregressive Distributed Lag/Error Correction Model can be utilised to create a correctly specified model where standard inference and hypothesis testing applies. This would be a possible extension of the work here but was not attempted.

Nevertheless, cointegration was checked to make sure spurious regression is not what we found. The presence of cointegration further strengthens the argument that the Relative Sentiment is related to macroeconomic and other comparative time series, at least in the long run, and therefore provide further strength to the Granger causality results. Johansen's and Philips-Ouliaris cointegration test was used. Please consult Johansen (1991) and Phillips and Ouliaris (1990) for more detail.

Table 18: Cointegration Tests

Cointegration Test: Phillips-Ouliaris					
Time Series	GDP, nom, QoQ, growth	GDP, real, QoQ, growth	VIX	UMCSENT	EPU
Relative Sentiment	-112.93^{***}	-110.73^{***}	–	–	–
Entropy	–	–	–	–	–
Cointegration Test: Johansen-Procedure					
Time Series	GDP, nom, QoQ, growth	GDP, real, QoQ, growth	VIX[59]	UMCSENT	EPU
Relative Sentiment	r = 0: 27.32^{***} r = 1: 6.52	r = 0: 28^{***} r = 1: 4.84	–	–	–
Entropy	–	–	–	–	–

Note: This table reports the results of cointegration tests according to Phillips and Ouliaris (1990) and Johansen (1991). The null hypothesis is that of *no* cointegration present. The Johansen-procedure tests explicitly for the amount of cointegration relationships in a VAR model. r^{max} is determined by the rank of the VAR model -1. Since I test only two series at one time in one model, $r^{max} = 1$. If $r = 0$ is rejected and $r = 1$, the series are cointegrated when both I(1). For Entropy, the statistic was insignificant at both $r = 0$ and $r = 1$. *p<0.1; **p<0.05; ***p<0.01

[59] The time series of the variable is stationary, and thus, per definition, cannot be cointegrated with another series. Cf. the stationarity tests in Tables 15 and 16.

Appendix J Time Series Comparisons

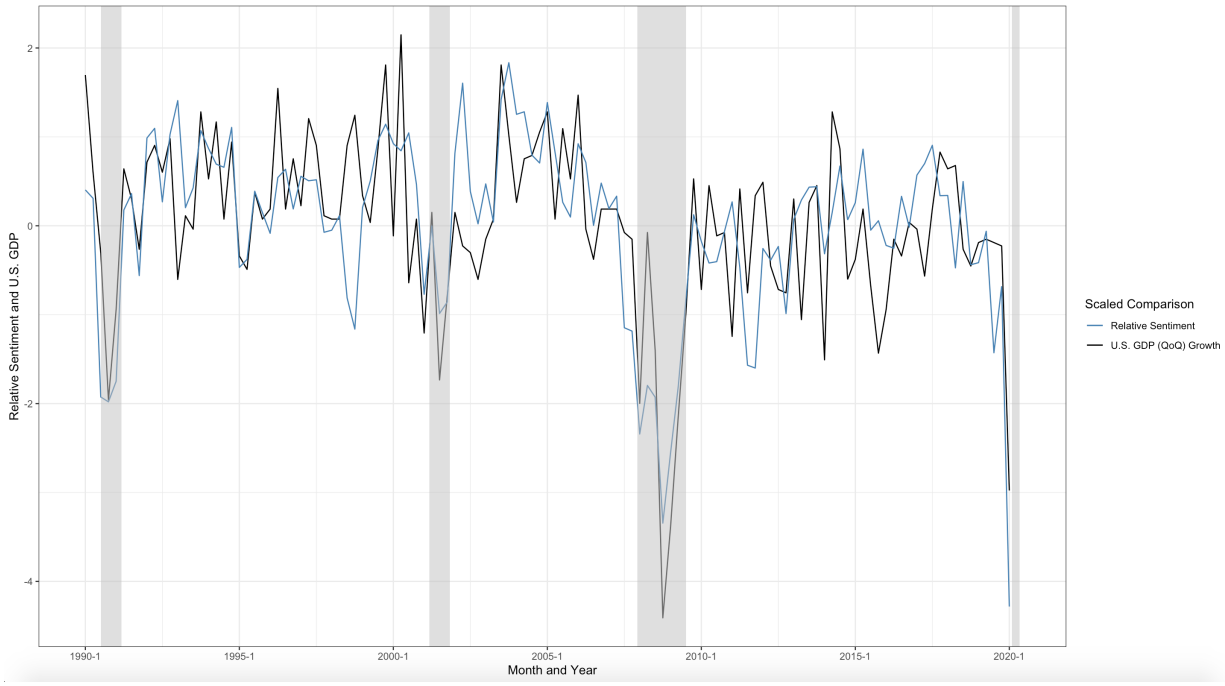


Figure 25: Scaled Comparison: U.S., QoQ Growth (nominal, seasonally adjusted) and Relative Sentiment

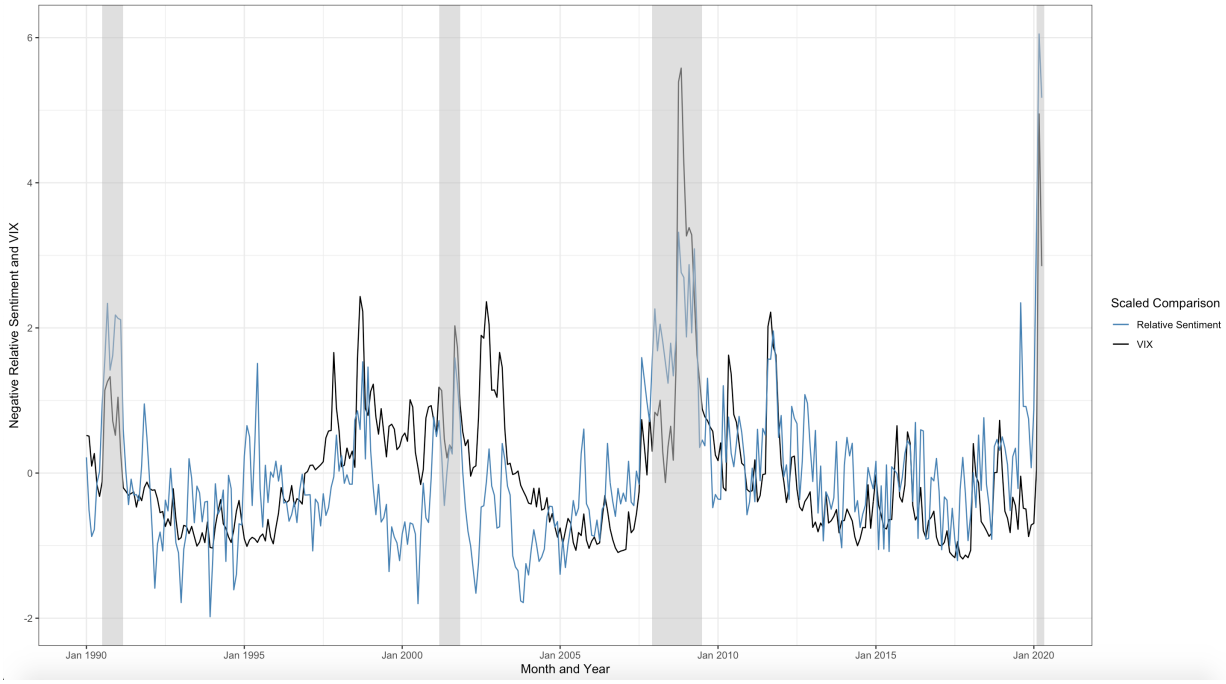


Figure 26: Scaled Comparison: CBOE Volatility Index and Relative Sentiment

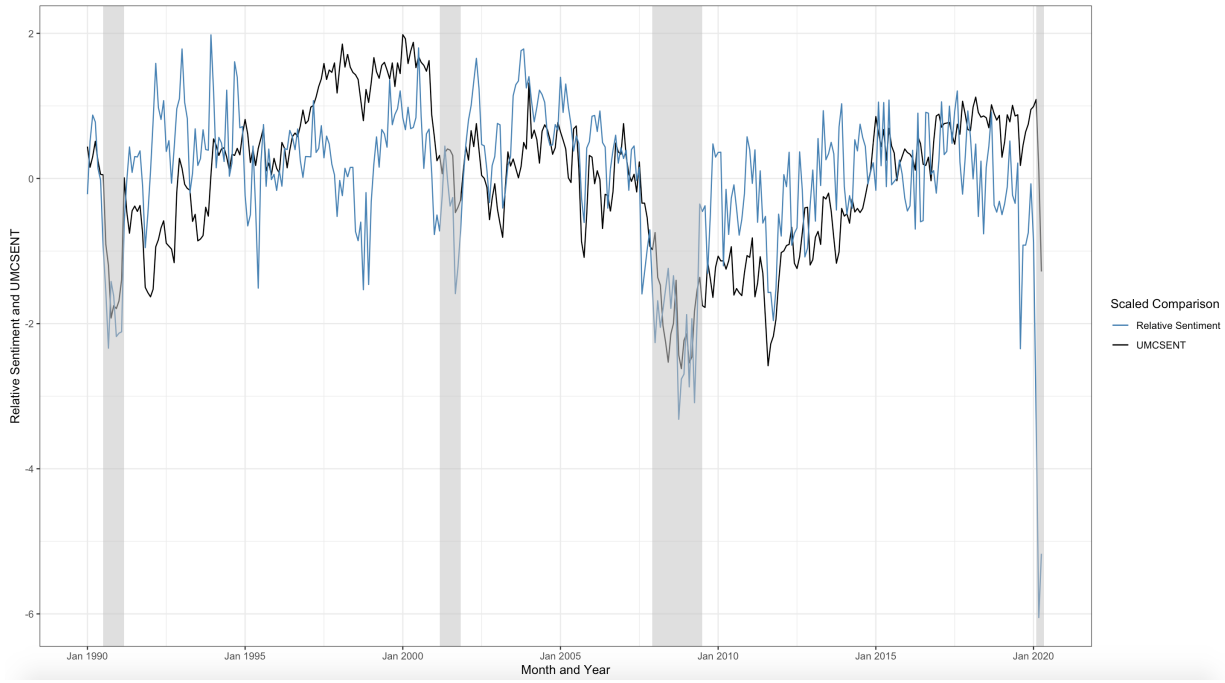


Figure 27: Scaled Comparison: University of Michigan Consumer Confidence Survey and Relative Sentiment

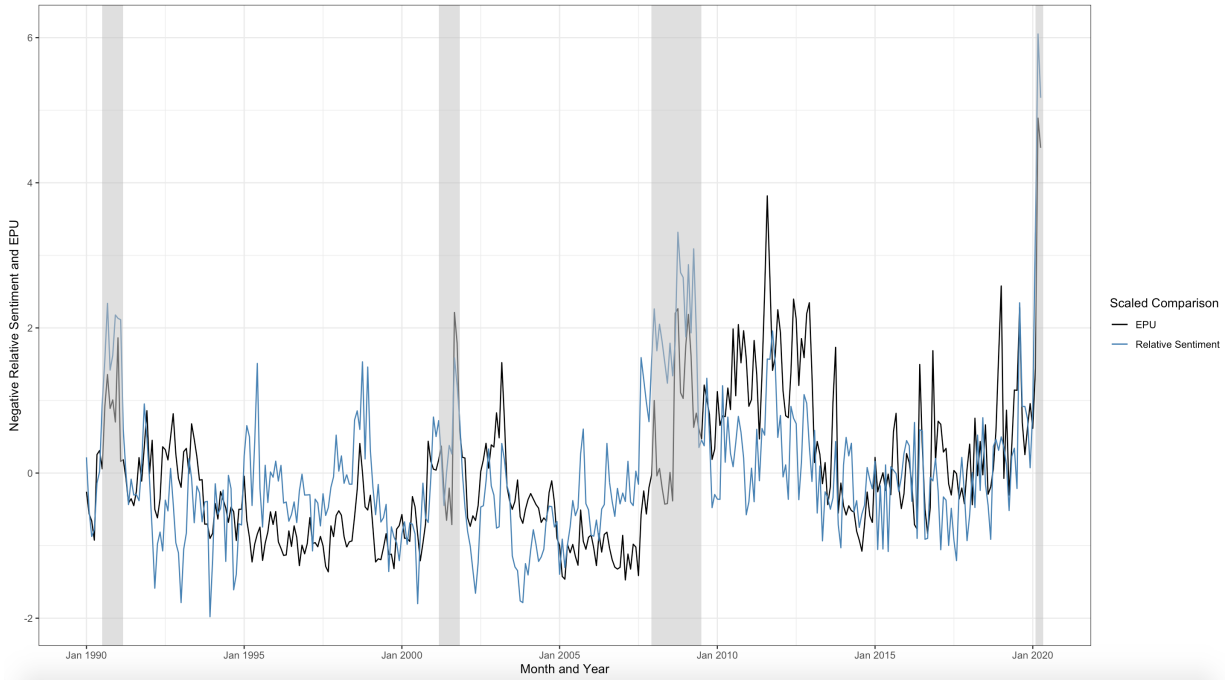


Figure 28: Scaled Comparison: Economic Policy Uncertainty Index (U.S.) and Relative Sentiment

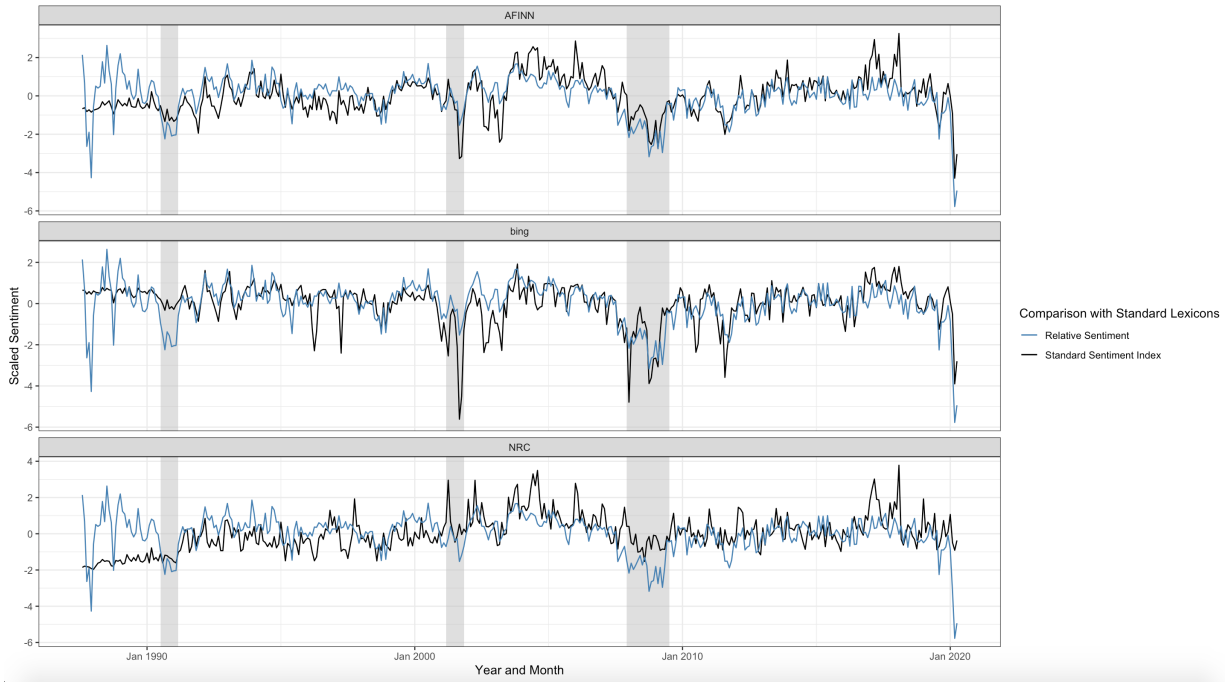


Figure 29: Scaled Relative Sentiment Index and Sentiment Scores Based on the Standard Lexical Sentiment Scores of Årup Nielsen (2011), Bing and Minqing (2004) and Mohammad and Turney (2013)

Appendix K LDA Topic Model

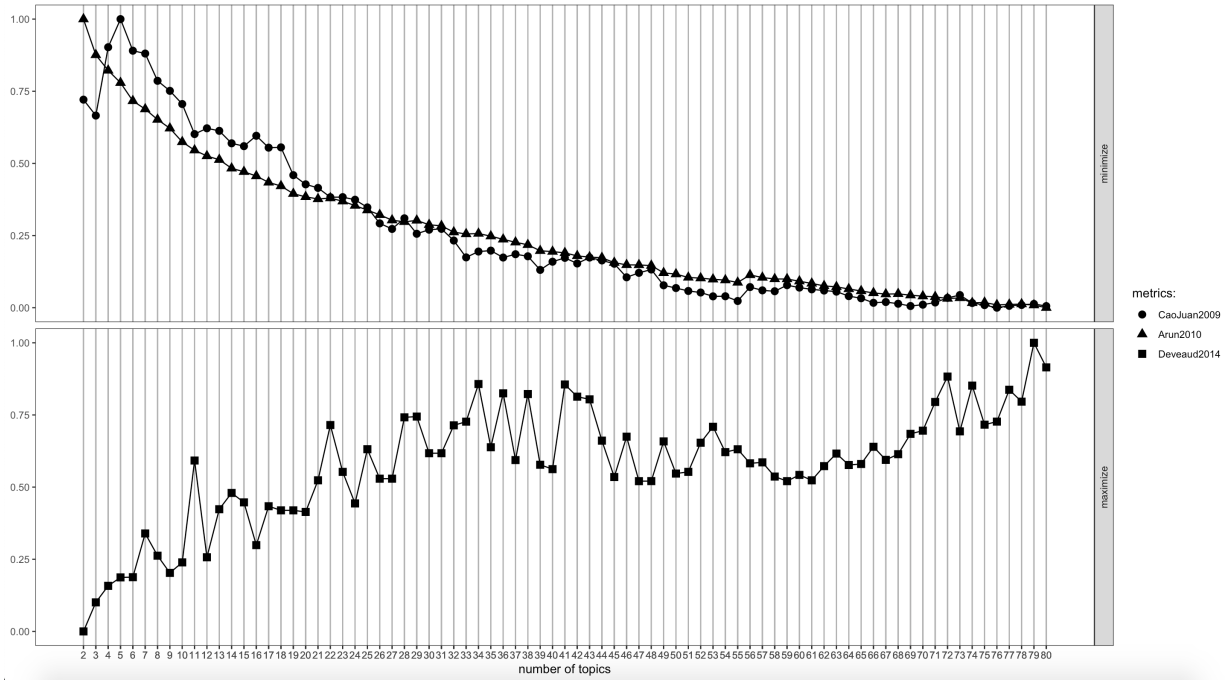


Figure 30: Choosing LDA Hyperparameter: Number of Topics. Based on Cao et al. (2009), Arun et al. (2010) and Deveaud et al. (2014).

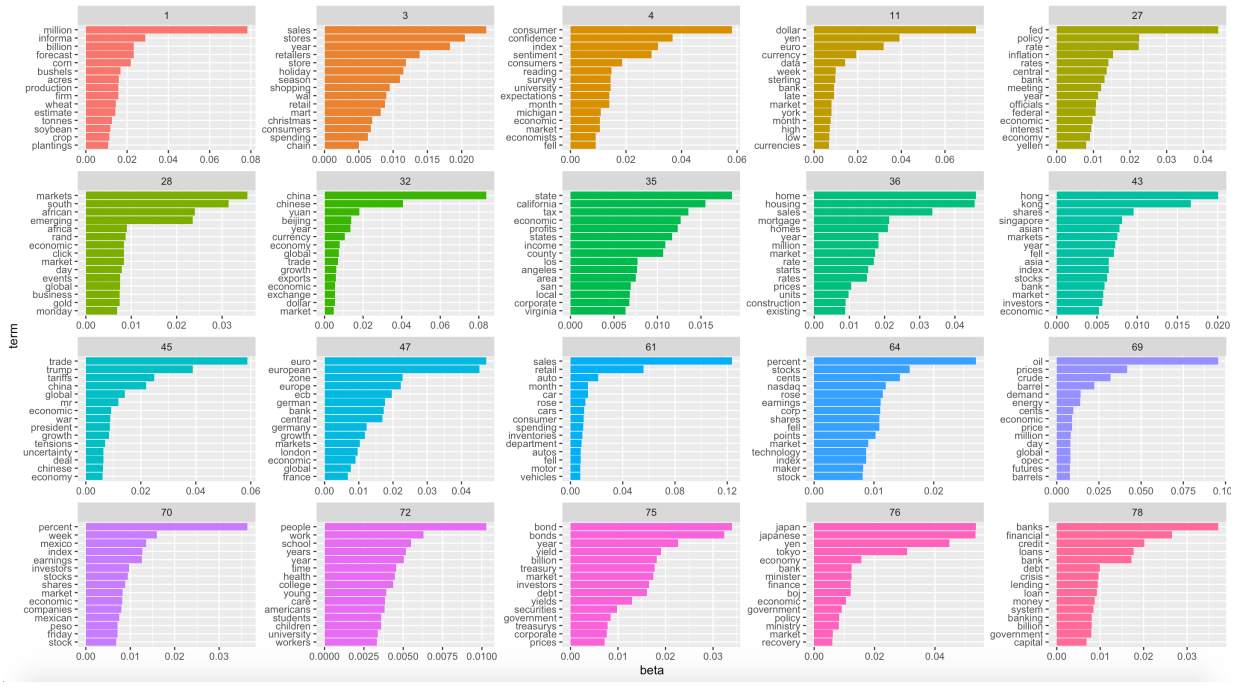


Figure 31: LDA Model: Topics. Choice of Interesting Ones. Topics Represented by Top 15 Words. Possible Interpretation: 1 – Agriculture; 3 – Retail Sales; 4 – University of Michigan Consumer Confidence Survey; 11 – Forex Markets; 27 – FED (under Yellen); – 28 – FED (under Greenspan); 32 – China and Chinese Trade; 35 – California; 36 – Housing Market; 43 – Asian Economy and Markets; 45 – Trade War with China; 47 – European Union and ECB; 61 – Car Sales; 64 – Technology Stocks; 69 – Oil; 70 – Mexico; 72 – Students and Youth; 75 – Bond Markets; 76 – Japan; 78 – Bank Lending.

Appendix L Entropy: Index Cross-Correlation Function

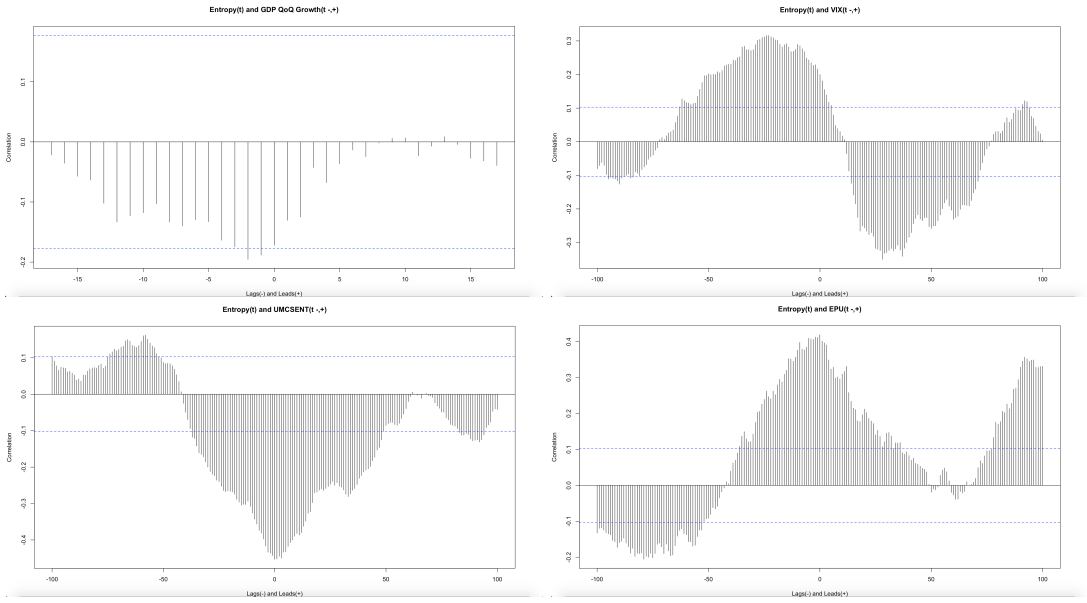


Figure 32: Cross-Correlation Functions. $Entropy_t$ correlated with $Comparison_s$ where $s = (-15, ..., 15)$ in the first graph and $s = (-100, ..., 100)$ in graphs 2-4. *Comparison* refers to the four comparable time series: Nominal GDP quarter on quarter U.S. growth, CBOE Volatility Index, University of Michigan Consumer Sentiment Index and the Economic Policy Uncertainty Index. Correlation with $s > 1$ should be suggestive of predictive power.

Appendix M Entropy: Further Linear Regressions

Table 19: Entropy Index: Further Linear Regressions

	Dependent variable:			
	Nominal_GDP_QoQ_Growth			
	(1)	(2)	(3)	(4)
RS_L1			9,235.071***	9,209.707***
			(2,583.672)	(2,609.342)
RS_L2			21.913	−260.803
			(2,757.296)	(3,089.495)
RS_L3			427.792**	2,999.840
			(163.302)	(3,139.680)
RS_L4			−408.824**	−1,861.692
			(161.843)	(2,655.487)
RS_L5			292.841*	261.661
			(161.295)	(166.816)
RS_L6			−137.475	−152.677
			(130.653)	(132.460)
Entropy_L1	−2.617	−1.217	6.228**	6.848**
	(1.813)	(3.452)	(2.847)	(2.906)
Entropy_L2		−5.723	−4.584*	−5.512
		(4.499)	(2.739)	(3.590)
Entropy_L3		5.892		5.283
		(4.512)		(3.580)
Entropy_L4		−1.559		−5.189*
		(3.449)		(2.783)
RS_L1:Entropy_L1			−2,276.272***	−2,272.896***
			(698.871)	(706.369)
RS_L2:Entropy_L2			−43.168	50.639
			(743.007)	(837.038)
RS_L3:Entropy_L3				−691.049
				(850.519)
RS_L4:Entropy_L4				395.162
				(718.366)
Constant	13.683**	13.699*	−2.411	−1.680
	(6.614)	(7.509)	(6.589)	(7.077)
Observations	121	118	116	116
R^2	0.017	0.038	0.476	0.497
Adjusted R^2	0.009	0.004	0.426	0.427
Residual Std. Error	4.300 (df = 119)	4.338 (df = 113)	3.315 (df = 105)	3.312 (df = 101)
F Statistic	2.084 (df = 1; 119)	1.131 (df = 4; 113)	9.544*** (df = 10; 105)	7.125*** (df = 14; 101)
Note:				*p<0.1; **p<0.05; ***p<0.01

Appendix N Entropy: Structural Break Analysis

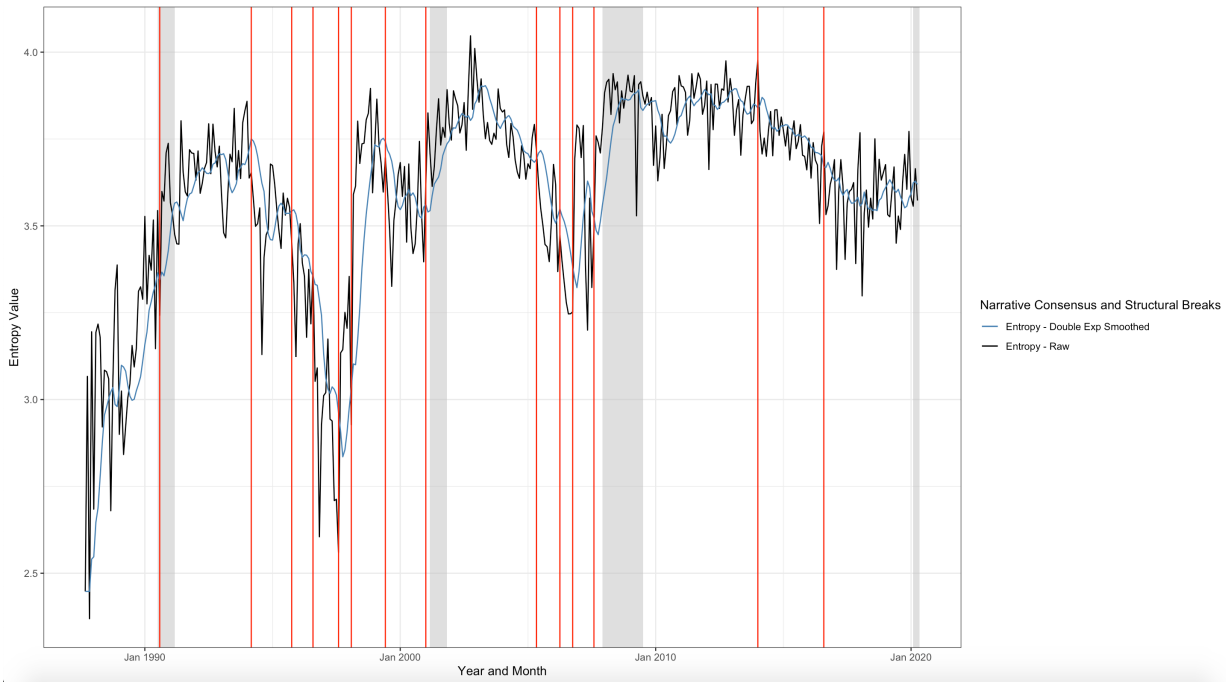


Figure 33: Structural Breaks in the Entropy Index: Monthly

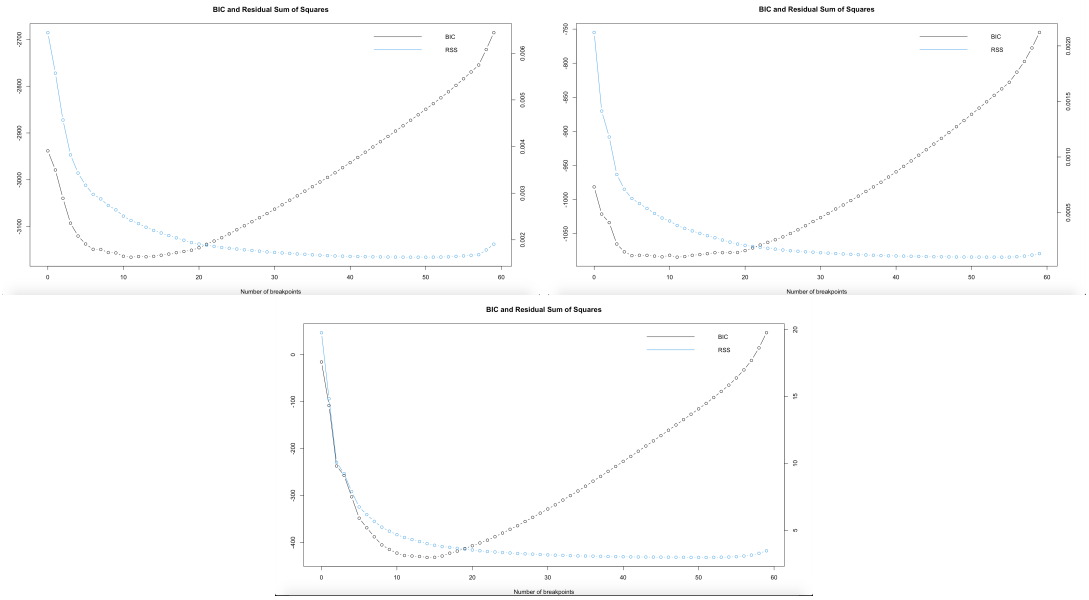


Figure 34: Structural Break Analysis: BIC Statistics based on Bai and Perron (2003). Top (left): Monthly RS_t Breaks. Top (right): Quarterly RS_t Breaks. Bottom: $Entropy_t$ Breaks.

Appendix O Word Embeddings: Robustness

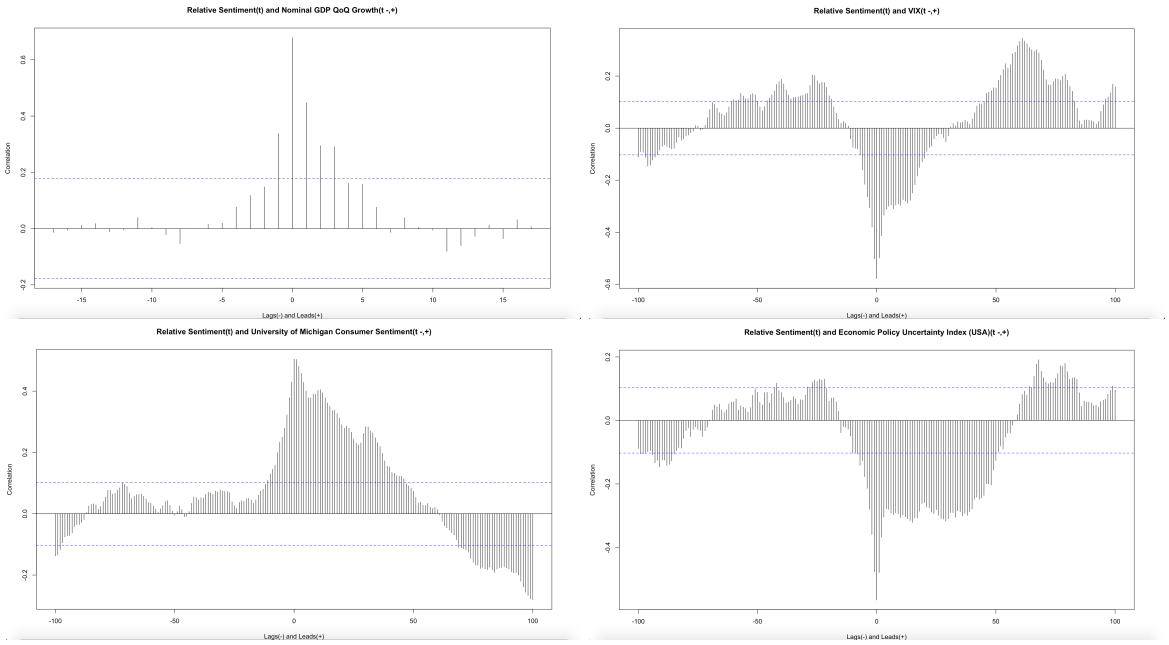


Figure 35: Cross-Correlation Functions. RS_t (Robust 1) correlated with $Comparison_s$ where $s = (-15, ..., 15)$ in the first graph and $s = (-100, ..., 100)$ in graphs 2-4. *Comparison* refers to the four comparable time series: Nominal GDP quarter on quarter U.S. growth, CBOE Volatility Index, University of Michigan Consumer Sentiment Index and the Economic Policy Uncertainty Index. Correlation with $s > 1$ should be suggestive of predictive power.

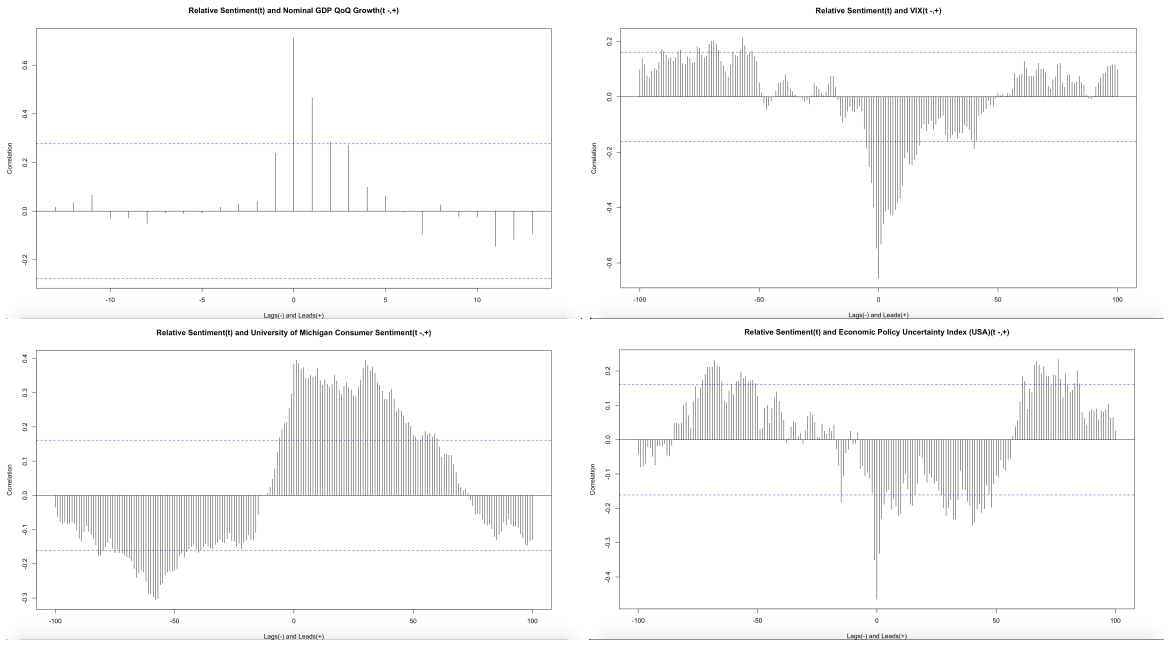


Figure 36: Cross-Correlation Functions. RS_t (Robust 2) correlated with $Comparison_s$ where $s = (-15, ..., 15)$ in the first graph and $s = (-100, ..., 100)$ in graphs 2-4. *Comparison* refers to the four comparable time series: Nominal GDP quarter on quarter U.S. growth, CBOE Volatility Index, University of Michigan Consumer Sentiment Index and the Economic Policy Uncertainty Index. Correlation with $s > 1$ should be suggestive of predictive power.

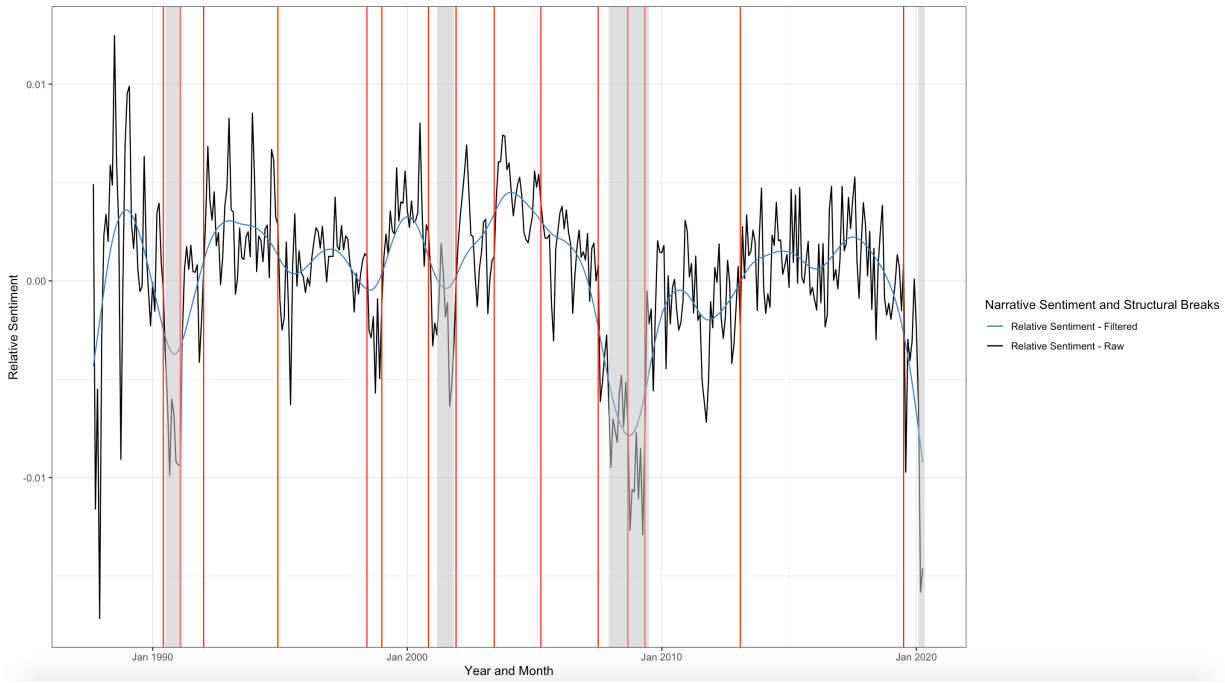


Figure 37: The Relative Sentiment Index (Robust 1). Estimated Structural Breaks Included. Word Embeddings Without 'Visionary' Terms

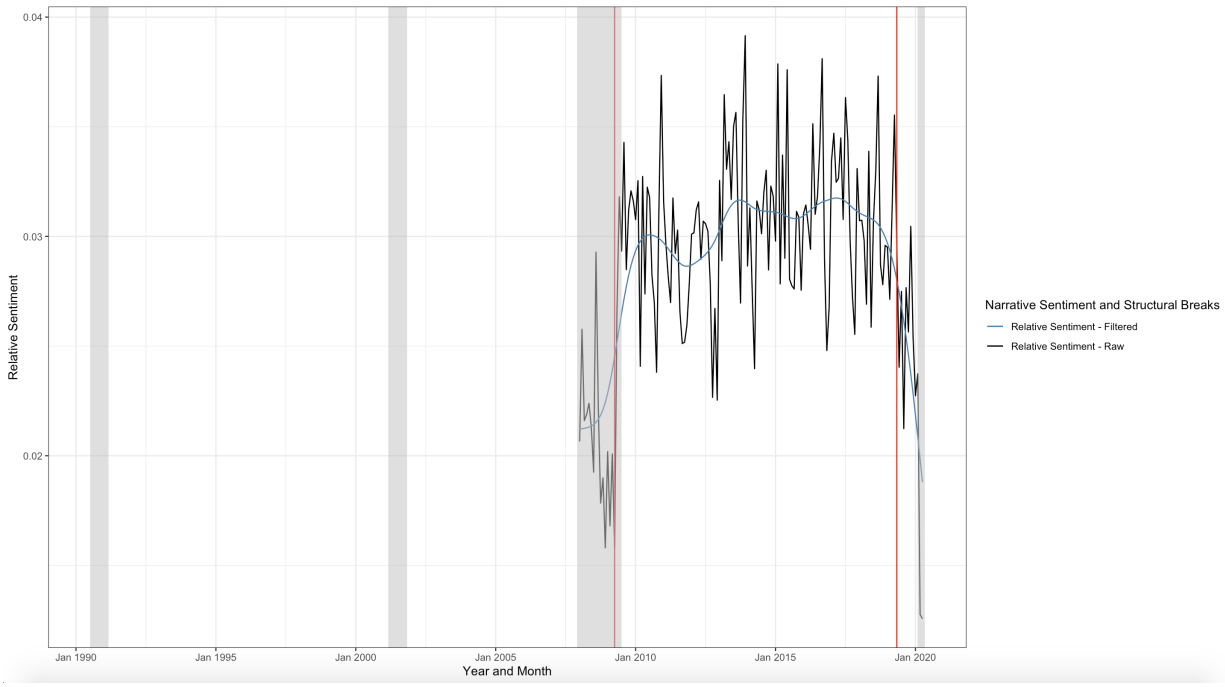


Figure 38: The Relative Sentiment Index (Robust 2). Estimated Structural Breaks Included. Word Embeddings Estimated With Pre-2008 News Articles

Appendix P Word Embeddings: Interesting Patterns

Figure 39: Vector Space Word Embedding Relationships. Red Used to Highlight Interesting Results

$$\mathbf{w}_{\text{subprime}} + \mathbf{w}_{\text{lending}} \approx$$

	word	cosinedistance
1	lending	0.85
2	subprime	0.84
3	mortgage	0.74
4	mortgages	0.71
5	loans	0.68
6	lenders	0.68
7	borrowers	0.68
8	mortgage_lending	0.66
9	subprime_mortgage	0.65
10	loan	0.58
11	bank_lending	0.58
12	banks	0.58
13	credit	0.57
14	risky	0.55
15	subprime_mortgages	0.54
16	mortgage_market	0.52
17	defaults	0.51
18	refinancing	0.51
19	mortgage_loans	0.51
20	commercial	0.50

$$\mathbf{w}_{\text{yellen}} - \mathbf{w}_{\text{fed}} + \mathbf{w}_{\text{ecb}} \approx$$

	word	cosinedistance
1	ms_yellen	0.66
2	janet_yellen	0.66
3	yellen	0.65
4	ms_merkel	0.64
5	trichet	0.62
6	claude_trichet	0.57
7	merkel	0.56
8	janet	0.54
9	claude	0.53
10	chairwoman	0.52
11	jean_claude	0.52
12	ecb	0.51
13	angela_merkel	0.49
14	jean	0.49
15	lucas	0.47
16	ms	0.45
17	mester	0.44
18	angela	0.44
19	mr_trichet	0.44
20	daly	0.43
21	mario_draghi	0.43

Appendix Q List of Resources

Table 20: List of Resources

All R Packages Used in the Master's Thesis		
R Package Name	Authors	Papers Introducing
tm	Ingo Feinerer, Kurt Hornik and David Meyer	Feinerer, Hornik, and Meyer (2008) and Feinerer and Hornik (2019)
tm.plugin.factiva	Milan Bouchet-Valat	Bouchet-Valat (2019)
topicmodels	Bettina Grün and Kurt Hornik	Grün and Hornik (2011)
ldatuning	Nikita Murzintcev	Murzintcev (2020)
text2vec	Dmitriy Selivanov, Manuel Bickel and Qing Wang	Selivanov et al. (2020)
forecast	Rob J Hyndman and Yeasmin Khandakar	Hyndman and Khandakar (2008)
stats	R Core Team	R Core Team (2020)
urca	Bernhard Pfaff	Pfaff (2008a)
tidytext	Julia Silge and David Robinson	Silge and Robinson (2016)
tidyverse	Numerous Authors	Wickham et al. (2019)
zoo	Achim Zeileis and Gabor Grothendieck	Zeileis and Grothendieck (2005)
lubridate	Garrett Grolemund and Hadley Wickham	Grolemund and Wickham (2011)
vars	Bernhard Pfaff	Pfaff (2008b)
aod	M. Lesnoff and R. Lancelot	Lesnoff and Lancelot (2012)
aTSA	Debin Qiu	Qiu (2015)
ggplot	Hadley Wickham	Wickham (2016)
stargazer	Marek Hlavac	Hlavac (2018)

Additional Resources Used in the Master's Thesis	
Resource Name	Sources
Dow Jones Factiva	*Factiva* (n.d.)
Federal Reserve Bank of St. Louis Economic Data (FRED)	Baker et al. (n.d.); Chicago Board Options Exchange (n.d.); University of Michigan (n.d.); U.S. Bureau of Economic Analysis (n.d.-a, n.d.-b)
National Bureau of Economic Research	National Bureau of Economic Research (n.d.)