The impact of news narrative on the economy and financial markets

Sonja Tilly

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

May 27, 2022

I, Sonja Tilly, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis investigates the impact of news narrative on socio-economic systems across four experiments.

Recent years have witnessed a rise in the use of so-called alternative data sources to model and predict dynamics in socio-economic systems. Notably, sources such as newspaper text allow researchers to quantify the elusive concept of narrative, to incorporate text-based features into forecasting frameworks and thus to evaluate the impact of narrative on economic events.

The first experiment proposes a new method of incorporating a wide array of sentiment scores from global newspaper articles into macroeconomic forecasts, attempting to forecast industrial production and consumer prices leveraging narrative and sentiment from global newspapers. I model industrial production and consumer prices across a diverse range of economies using an autoregressive framework.

The second experiment uses narrative from global newspapers to construct theme-based knowledge graphs about world events, demonstrating that features extracted from such graphs improve forecasts of industrial production in three large economies.

The third experiment proposes a novel method of including news themes and their associated sentiment into predictions of changes in breakeven inflation rates (BEIR) for eight diverse economies with mature fixed income markets. I utilise five types of machine learning algorithms incorporating narrative-based features for each economy.

In the above experiments, models incorporating narrative-based features generally outperform their benchmarks that do not contain such variables, demonstrating the predictive power of features derived from news narrative.

The fourth experiment utilises GDELT data and the filtering methodology introduced in the first experiment to create a profitable systematic trading strategy based on the average tone scores for 15 diverse economies.

Impact Statement

The findings in this thesis demonstrate that news narrative has a significant impact on socio-economic systems. Features extracted from news stories improve forecasts of major macroeconomic indicators for a diverse set of countries. Further, such narrative-based features can be leveraged to track in real time, and to potentially identify at an early stage, specific socio-economic phenomena.

These results are relevant to institutions such as central banks that monitor the state of the economy and base wide-reaching policy decisions on their forecasts.

Further, the findings in this thesis are pertinent to the work of supranational organisations such as the United Nations whose bodies monitor phenomena like migration, civil unrest and climate change.

This research is of potential interest to the asset management industry where features from news topic-based knowledge graphs can help to early identify regime changes and thus provide valuable support for asset allocation decisions. Further, signals extracted from narrative can be transformed into trading strategies that can diversify, and potentially enhance, traditional sources of alpha.

Within academia, this research contributes to the literature on enhancing macroeconomic forecasts with narrative-based features. Moreover, the thesis advances literature on economic knowledge graphs and proposes an analytical framework that sets out how news topic-based graphs can be used to monitor a multitude of socio-economic phenomena in real-time.

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance.

I would first like to thank my primary supervisor, Dr. Giacomo Livan, whose expertise and friendly guidance was invaluable in formulating the research questions and methodology. Your insightful feedback helped me to sharpen my thinking and brought my work to a higher level.

I would also like to thank my secondary supervisor Prof. Dr. David Tuckett, for his valuable feedback throughout my studies.

I would like to acknowledge my colleagues at Quoniam Asset Management for their friendly collaboration. I would particularly like to mention my supervisor at Quoniam Asset Management, Dr. Markus Ebner. Markus, I want to thank you for your patient support and for the opportunity to further my research in industry.

In addition, I would like to thank my husband Joel for his loving support, sympathetic ear and for always being there for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stories are an inherent part of human society and culture and their dynamics and impact on socio-economic systems – particularly in the run-up to regime changes or events of systemic importance – are well-documented (Keynes, 1937; Shiller, 2019).

Given the real-time nature of global news reporting, the narrative emerging from the news may carry information that is not yet reflected in such systems.

A growing number of works attempt to predict the economy and financial markets utilising so-called “alternative data“ such as text in addition to conventional data. However, the large majority of literature in this field focuses on large economies such as the US and anglophone sources of narrative, often from one single news outlet. Hence, most research in this field ignores the multi-faceted nature of global news narrative, offering a lot of potential for further research and investigation.

The Global Database of Events, Language and Tone (GDELT) – the database used in this thesis – monitors global media reporting from a multitude of different perspectives and extracts entities from raw news text, such as themes, emotions, locations and many more. While being a very rich data source and freely accessible on the Google Cloud Platform, GDELT is underutilised both in academia and industry, probably due to its non-commercial nature. For instance, GDELT does not offer user support and requires a good level of time and technical expertise to filter and transform the data into meaningful signals. This may result in GDELT being less accessible to a broad user base compared to commercial offerings such as Ravenpack, Refinitiv or S&P Global.

Developments in technology and the evolution of powerful computational infrastructure allow to quantify the concept of news narrative as extracted from GDELT data and to extract features for inclusion into analytical and predictive frameworks. This thesis brings together concepts from different fields such as economics, psychology, network science and computer science and applies them to examine global news narrative in a domain that offers plenty of scope for further academic research. Parts of this research were conducted in collaboration with Quoniam Asset Management under the supervision of Dr. Markus Ebert, providing expert guidance and confirming the industry demand for this work.

1.1 Research objectives

The main objective of this thesis is to examine and to quantify the impact of news narrative on socio-economic systems leveraging GDELT as data source. To address this objective, four experiments have been carried out to operationalise concepts such as information extraction and feature generation based on GDELT, to investigate the predictive power of narrative-based features for a range of economic and market-based indicators across a diverse set of geographies and to analyse the key performance drivers of such predictions.

1.2 Scientific contributions

This thesis advances existing research in a number of aspects:

- Extracting informative signals from GDELT, a very rich yet underused data source. I develop a machine learning-based filtering and feature extraction methodology for GDELT and demonstrate its effectiveness in distilling signals from vast volumes of data.
- Quantifying the impact of news narrative on socio-economic systems. I

predict macroeconomic and market-based indicators leveraging news-based features and demonstrate that they generally improve forecasts compared to benchmarks that do not include these features.

- Identifying key performance drivers. I supplement my results with an interpretability study that identifies the key performance drivers from the narrative-based factors incorporated into predictive frameworks.

- Leveraging graph theory and newspaper themes to monitor socio-economic phenomena. I use theme-based graphs to show that phenomena such as the COVID-19 outbreak can be tracked, and possibly identified early, in real-time. Further, I examine cross-country spillover effects of news reporting on breakeven inflation rates (BEIR).

- The contents of chapter 4 was published as Tilly, S., Ebner M. and Livan G. (2021). Macroeconomic forecasting through news, emotions and narrative. Expert Systems with Applications, 175:114760.

- The contents of chapter 5 was published as Tilly, S. and Livan G. (2021). Macroeconomic forecasting with statistically validated knowledge graphs. Expert Systems with Applications, 186:115765.

- The contents of chapter 6 was submitted as an article to a scientific journal and is currently under review. The manuscript is available as Tilly, S. and Livan, G. (2021). Predicting market inflation expectations with news topics and sentiment. https://arxiv.org/pdf/2107.07155.pdf.

1.3 Thesis outline

This thesis is structured as follows:

- Chapter 2 presents a choice of relevant background literature on the main concepts covered in this research. It discusses the role of narrative in economic cycles, addresses the importance of sentiment and narrative in decision-making and discusses the operationalisation of narrative from a Computer Science perspective.

- Chaper 3 outlines the foundation of the statistical methodologies used in this thesis. It sets out the analysis problems and the solutions proposed by existing methods and justifies the methodological choices in this research.

- Chapter 4 introduces a new approach of including sentiment from news reporting into macroeconomic forecasts, looking to predict industrial production and consumer prices. It demonstrates the effectiveness of the thematic data filtering methodology proposed by comparing model performance for filtered and unfiltered data. Results are supplemented by an interpretability study that identifies the main performance drivers associated with the original GDELT emotion scores.

- Chapter 5 uses narrative from global newspapers to build theme-based knowledge graphs about world events, showing that features derived from such graphs enhance forecasts of industrial production in three geographies compared to a number of benchmarks. An interpretability study outlines those theme categories that represent the key performance drivers.

- Chapter 6 proposes a novel method of including news themes and their associated sentiment into predictions of breakeven inflation rate (BEIR) movements for eight economies with mature fixed income markets. A feature importance analysis sets out the main performance drivers in terms of topic groups. Further, the chapter contains an analysis of the cross-country spillover effects of news reporting on BEIR.

- Chapter 7 presents a news sentiment-based systematic trading strategy leveraging the filtering methodology introduced in chapter 4.

- Chapter 8 concludes this research, summarises its scientific contributions and outlines future projects to be carried on by other researchers.

Chapter 2

Background Literature

This chapter presents a selection of background literature that is of relevance to the dissertation as a whole, while each of the experimental chapters reviews topic-specific literature. Some of the literature reviewed in this chapter belongs to domains such as Psychology and Economics. Theories relating to narrative often stem from Social Sciences, while their operationalisation is developed in Computer Science. The first section discusses the role of narrative in economic cycles. The second section addresses the importance of sentiment and narrative in decision-making and discusses the operationalisation of the concept of narrative from a Computer Science perspective.

2.1 Narrative and economic cycles

Research on the impact of narrative and sentiment on the economy and financial markets has a long history and spans disciplines such as Psychology, Cognitive Sciences, Economics and Computer Science (Bruner, 1990; Brosch et al., 2013; Shiller, 2019). Economic events such as bubbles or crashes follow established patterns where narrative is an important driver, for instance in the build-up of irrational exuberance (Reinhart and Rogoff, 2009; Aliber and Kindleberger, 2017). In his classic work on Economics, Keynes states that emotional factors – so-called “animal spirits” – determine human actions and have quantifiable economic effects

(Keynes, 1937). Akerlof and Shiller build on the concept of animal spirits, suggesting that emotional factors must be taken into account to satisfactorily interpret the drivers of economic events (Akerlof and Shiller, 2010). The diagnostic expectations theory states that market participants react excessively to news and subsequently develop distorted expectations of the likelihood of economic outcomes (Gennaioli and Shleifer, 2018). Along with the idea of animal spirits, the concept of diagnostic expectations helps explain investor behaviour and events in financial markets beyond neoclassical economic theory. In his book "Narrative Economics", Shiller draws an analogy between narrative and epidemiology when describing how narratives spread rapidly before eventually slowing down, with causality between narrative and economic outcomes existing both ways (Shiller, 2019).

2.2 Narrative, decision-making and their operationalisation

Research on the impact of narrative on socio-economic systems is an interdisciplinary field, with theories developed in Social Sciences and most operationalisations developed by Computer- or Data Scientists.

Existing research in Psychology and Cognitive Sciences finds that emotions support making advantageous choices, particularly in complex situations with uncertain results (Damasio, 1996; Brosch et al., 2013). Indeed, positive (negative) emotions promote (inhibit) individual actions (Clore and Palmer, 2009). Narrative helps individuals make sense of the unusual and, on a collective level, is used to develop culture (Bruner, 1990). Conviction Narrative Theory (CNT) considers narratives as a means for agents to envision future outcomes, with agents seeking reassurance in narrative, particularly when faced with uncertainty (Tuckett et al., 2014). Extending the concept of CNT, Nyman *et al.* find that fluctuations in news sentiment precede changes in the economy and therefore are predictive of economic indicators and financial market movements (Nyman et al., 2021).

Newspapers are an established means to circulate information and thus, a major channel for news reporting. Today, most newspapers have an online presence and generate large amounts of text. This text incorporates information in the form of viewpoints and sentiment about the economy or financial markets, which may not yet be reflected in macroeconomic indicators or market prices.

Recent developments in technology, cloud infrastructure and natural language processing permit to analyse news narrative and its impact on socio-economic systems on a big scale. This has led to the inclusion of text-based features – alongside conventional variables – into econometric models such as vector autoregressions (Stock and Watson, 2001), or structural frameworks such as dynamic stochastic general equilibrium models (Christiano et al., 2005; Smets and Wouters, 2007).

Newspapers with an online presence as well as social media platforms produce huge volumes of textual data on an ongoing basis. Therefore, the extraction of text-based features is typically a task that involves processing "big data".

Doornik and Hendry group big data into three main types (Doornik and Hendry, 2015):

- Tall: many observations *T*, fewer variables *N*, with $T \gg N$. Examples include tick-by-tick data of financial transactions.

- Fat: many variables *N*, fewer observations *T*, with $N \gg T$. Such information includes large cross-sectional data sets.

- Huge: both large number of variables *N* and observations *T*. Such data sets are ideal for forecasting purposes, containing a wealth of information as well as sufficient track record for cross-validation. This is the typical case when dealing with news data and applies to GDELT, the data set used in this thesis.

Big data collection has started recently, typically limiting the historical track records to below a decade.

The United Nations Economic Commission for Europe (UNECE) propose the following taxonomy for classifying big data types (UNECE, 2021):

- Human-generated information documenting human experiences as text, pictures, audio and video and is loosely structured. Examples include social media platforms such as Twitter, Instagram, TikTok, YouTube and GDELT, which is the data set used in this research.

- Data from business processes recording events such as registering a customer, manufacturing a product or taking an order, typically well-structured. Examples of such data are commercial transactions, credit card data and medical records.

- Machine-generated data, extracted from sensors and machines employed to measure and record events and situations in the physical world, for instance changes in weather, pollution levels, traffic or satellite images.

In terms of big data types, this thesis focuses on the first category described above, namely the analysis of newspaper narrative.

There is a growing body of literature examining sentiment from different kinds of media and its predictive power in relation to the economy and financial markets.

Rousidis reviews the use of large volumes of social media narrative for predictions in a wide range of areas such as finance, marketing and sociopolitics (Rousidis et al., 2020). Findings indicate that the inclusion of narrative generally shows improved predictions in financial applications (stock markets, house prices) but proves less reliable in sociopolitical (elections) and marketing domains. Literature in the finance domain uses media sentiment prediction either for specific assets or sectors (micro level) (Allen et al., 2019) or for macroeconomic indicators such as GDP (macro level) (Ardia et al., 2019). Some works explore large volumes of unstructured text from various media types to generate signals (Buono et al., 2018; Elshendy et al., 2018). Other papers describe methods to include such signals into forecasting frameworks, for example to better monitor the economy and financial markets (Levenberg et al., 2014; Slaper et al., 2018).

According to a study by Kapetanios and Papailias, the inclusion of features derived from big data sets such as Google Search data improves predictions of economic variables due to their real-time nature (Kapetanios and Papailias, 2018). The authors highlight the need of applying features selection or dimensionality reduction techniques to isolate signals and point out that depending on the data source, access and track record may be limited. Buono *et al.* find that large scale data such as electronic payment data, scanner/online prices, online searches, textual data and social media narrative (Twitter) improve precision in economic forecasting (Buono et al., 2017). However, the authors caution that big data should be regarded as a complement to traditional explanatory variables, not a replacement.

Application-specific literature will be reviewed at the beginning of each of the four experimental chapters.

Chapter 3

Methodology and algorithms

This chapter provides the foundation of the statistical methodologies used in this thesis. It discusses the analysis problems and the solutions proposed by existing approaches and justifies the methodological choices in this research. The chapter starts by reviewing methods for testing time series for stationarity, followed by methods for causality analysis and for feature selection. It sets out the algorithms used in this research, methods for model performance evaluation and prediction comparison and ends with discussing knowledge graph principles and metrics.

3.1 Testing for time series stationarity

The statistical properties of a stationary time series do not vary with time. Economic and market-based time series often show trending behaviour or exhibit non-stationarity in the mean, requiring some form of trend removal such as first differencing. The following section reviews a selection of existing statistical tests to check for stationarity in time series.

The Augmented Dickey-Fuller is based on a test regression given by

$$y_t = \beta' D_t + \phi y_{t-1} + \sum_{j=1}^{p} \psi_j \triangle y_{t-j} + \varepsilon_t \tag{3.1}$$

where D_t is a vector of deterministic terms (constant, trend, etc), ϕ is the coefficient of the first lag on y, the p-lagged difference terms $\triangle y_{t-j}$ represent a proxy for the autoregressive moving average structure of errors and the value of p is set so error ε_t is serially uncorrelated. ε_t is assumed to be homoskedastic. This test checks the null hypothesis that $\phi = 1$, i.e. that a unit root is present and the time series is therefore non-stationary (Dickey and Fuller, 1979). The Augmented Dickey-Fuller test addresses a key shortcoming of the original Dickey-Fuller test as it allows for higher-order autoregressive processes by incorporating lags of the order p, thus making it more thorough. This requires determining the lag length p when applying the test, which can be done by using the Bayesian Information Criterion (BIC) (see section 3.5.2.1 for details).

The $\mathrm{ADF_t}$ statistic is based on least squares estimates of Eq. (3.1) and is defined as

$$ADF_t = \frac{\hat{\phi} - 1}{SE(\phi)} \tag{3.2}$$

If the test statistic is larger than a set critical value, the null hypothesis cannot be rejected, indicating that the time series is non-stationary.

The Philips-Perron test is an alternative to the Augmented Dickey–Fuller and tests the null hypothesis that a time series has a unit root and is thus non-stationary (Phillips and Perron, 1988). Like the Augmented Dickey–Fuller test, the Phillips–Perron test tackles the problem that the autoregressive process may have a higher order of autocorrelation than is accounted for in the test equation. While the Augmented Dickey–Fuller test deals with this problem by introducing lags as regressors in the test equation, the Phillips–Perron test performs a non-parametric adjustment to the test statistic. According to Davidson *et al*, the Augmented Dickey-Fuller test and Philips-Perron test are asymptotically equivalent, however the Phillips–Perron test tends to underperform the Augmented Dickey–Fuller test in finite samples (Davidson et al., 2004).

In contrast to the Augmented Dickey-Fuller test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used for testing a null hypothesis that a time series is stationary around a deterministic trend (i.e. trend-stationary) and the alternative hypothesis states that it is non-stationary (Baum, 2018). Hence, in this particular test, the existence of a unit root is not the null hypothesis but the alternative. It is important to note that in this test, the absence of a unit root indicates trend-stationarity rather than stationarity. Therefore, using the KPSS test as an alternative to the Augmented Dickey-Fuller test may result in a misinterpretation of a time series' stationarity. Instead, the KPSS test supplements the augmented Dickey–Fuller test. In this research, the Augmented Dickey-Fuller test is used to test time series data for stationarity (see chapters 4, 5, 6 and 7) due to its advantages over the other methods considered in this section.

3.2 Causality analysis

This section discusses several approaches for testing if one time series is predictive of another.

The methods considered are both based on the idea proposed by Wiener that, given two interdependent variables x and y, y causes x if y improves predictions of future values of x, with x already predicting its own future (Wiener, 1956).

Granger Causality is a regression-based interpretation of the Wiener causality and was developed in econometrics. It is based on the concept that, if a time series of variable $x_i(t)$ "Granger-causes" a time series of variable $x_j(t)$, then past values of $x_i(t)$ include information that add value to the predictions of $x_j(t)$ over and above the information from past values of only $x_j(t)$ (Granger, 1969). $x_j(t)$ can be modelled in an autoregressive framework only using past values of itself

$$x_j(t) = \sum_{r=1}^{p} B_j x_j(t-r) + \varepsilon_j(t) \tag{3.3}$$

where p is the model order, B is the autoregressive coefficient and $\varepsilon_j(t)$ is the residual.

Equation (3.4) models $x_j(t)$ with past values of both $x_j(t)$ and $x_i(t)$

$$x_j(t) = \sum_{r=1}^{p} [A_{ji,r} x_i(t-r) + A_{jj,r} x_j(t-r)] + \varepsilon_{j|i}(t) \tag{3.4}$$

where p stands the the model order, matrices A incorporate the model coefficients and $\varepsilon_{j|i}(t)$ represents the residual. B_j, A_{ji} and A_{jj} can be found by least squares approach. The model order can be determined by using an information criteria such as BIC (see 3.5.2.1 for details).

If the variance of $\varepsilon_{j|i}(t)$ is less than $\varepsilon_j(t)$ by incorporating $x_i(t)$, then one can conclude that $x_i(t)$ Granger-causes $x_j(t)$. Therefore, Granger Causality can be quantified as

$$F_{x_i} \to x_j = \ln \frac{var(\varepsilon_j)}{var(\varepsilon_{j|i})} \tag{3.5}$$

An F-test can be used to test the hypothesis that x_i does not Granger-cause x_j. The Granger Causality test requires data stationarity and only considers linear relationships.

The more recent concept of Transfer Entropy is a measure of directed information transfer between joint processes (Schreiber, 2000). Transfer Entropy is focused on the resolution of uncertainty, while Granger Causality is based on prediction. Transfer Entropy is not limited to linear interactions between variables; however it is computationally more expensive than Granger Causality. Further, Barnett demonstrates that for vector autoregressive processes, Transfer Entropy corresponds to Granger Causality (Barnett et al., 2009).

In this research, Granger Causality is the preferred method to test if one time series is predictive of another as it is applied to economic time series variables that are modelled with an autoregressive framework (see chapters 4, 5, 6 and 7). In addition, this method scales easily to multiple variables.

3.3 Controlling for multiple testing

Multiple testing refers to multiple statistical inferences that are performed repeatedly, which increases the occurrences of type I errors (i.e. false positives).

There are two main approaches for controlling type I errors. First, the family-wise error rate calculates the probability of coming to at least one false conclusion in a series of hypothesis tests, i.e. the probability of at least one type I error. The term ‘family‘ refers to a number of tests.

Second, the false discovery rate denotes the expected proportion of false positives. The multiple comparisons problem can be corrected by recalculating the probabilities of a statistical test that is used multiple times.

A popular family-wise error correction procedure is the Bonferroni correction (Bland and Altman, 1995). In this approach, p_i represents the p-value for a test H_i; there are m repeated tests. The Bonferroni procedure rejects H_i if $p_i \leq \frac{\alpha}{m}$. A drawback of this approach is that it corrects for type I errors while inflating the type II errors. Further, it tends to be too conservative.

One widely used method for adjusting for false positives is the Benjamini Hochberg procedure (Benjamini and Yekutieli, 2005). This method controls the false discovery rate at α level. It recalculates the p-values and ranks them in ascending order, describing them by their new rank i.e. P_1, ..., P_m. For a chosen false discovery rate α, the Benjamini Hochberg procedure identifies the largest i so that $P_i \leq (\frac{i}{m})\alpha$. Unlike with family-wise error correction methods, the false discovery correction methods are adaptive so that a number of false positives is interpreted in the context of the total number of discoveries. Therefore, this research uses the Benjamini Hochberg procedure to correct for false positives. The procedure is applied when conducting multiple tests for Granger Causality in chapters 4, 5, 6 and 7.

3.4 Feature selection

This research works with high dimensional data sets, i.e. data sets with a significantly higher number of features than observations. These features exhibit intercorrelations, which can result in overfitting when incorporated into a model. To address this, some feature selection method has to be applied.

There are two main approaches to feature selection. The first approach only retains the most pertinent features from the original data, while the second approach creates a new, smaller set of features.

An example for the first group of approaches is Forward Feature Selection. Forward Feature Selection starts training a model n times using each feature separately. The feature generating the best performance is selected as starting feature. This is repeated adding the one feature generating the best performance improvement until no significant performance improvement can be observed. The main disadvantage of this approach is its computational expensiveness.

The second group of approaches applies dimensionality reduction to the original data set. This involves compressing a large feature space into a smaller number of uncorrelated features, that describe the underlying properties of the original variables.

Partial Least Squares (PLS) is a supervised method that derives two sets of scores from features (referred to as X) and the predicted variable (referred to as Y), respectively (De Jong, 1993). These features are calculated sequentially and require centred and scaled data. PLS decomposes X and Y so that

$$X = TP^T + E \tag{3.6}$$

$$Y = UQ^T + F \tag{3.7}$$

where X represents an $n \times m$ matrix of predictive features and Y represents an $n \times p$ matrix of predicted variables. T and U are $n \times l$ matrices that are projections of X and projections of Y, respectively. P and Q stand for, respectively, $m \times l$ and $p \times l$ orthogonal loading matrices. Matrices E and F represent the residuals, assumed to be independent and identically distributed random normal variables. PLS components are selected to maximise the covariance between T and U using eigenvalue decomposition. Each X-score represents a linear combination of X. For instance, the first X-score is given by $t = Xw$, where w is the eigenvector associated with the first eigenvalue of X^TYY^TX. Likewise, the first Y-score is given by $u = Yc$ where c is the eigenvector corresponding to the first eigenvalue of Y^TXX^TY. X^TY stands for the covariance of X and Y. After extracting the first component, the original X and Y values are deflated like so $X_1 = X - tt^TX$; $Y_1 = Y - tt^TY$ to remove the variability already explained, thus ensuring the orthogonality of components. The procedure is then repeated for the second component.

Cross-validation analysis can be applied to identify the optimal number of PLS components to extract. Linear regression models can be used to predict Y, each including an increasing number of PLS components. The model with the smallest residual sum of squares and its associated number of PLS components is appropriate (Tobias, 1995).

Another widely used dimensionality reduction technique, Principal Component Analysis (PCA), sequentially computes components that are linear combinations of X from centred and scaled data. In contrast to PLS, PCA looks to best explain feature space X and chooses scores T and loadings P so that each PCA component captures the largest variance of X. PCA can thus be considered an unsupervised approach.

Both PLS and PCA are suitable for extracting components from "fat" data sets (number of features » number of observations) where multicollinearity is an issue (Cubadda and Guardabascio, 2012). A key weakness of PCA is the fact that its components only capture characteristics of X, ignoring any relationship of X and Y. In tasks where a predicted variable Y is given, PLS is a more efficient dimensionality reduction technique than PCA due to its supervised nature. This means that PLS requires fewer components than PCA to achieve the same level of model performance (Maitra and Yan, 2008). Therefore, PLS is the preferred feature selection technique used in this research, applied in chapters 4, 5 and 6.

3.5 Algorithms

Machine learning algorithms can be categorised into three groups: supervised, unsupervised and reinforcement.

In supervised learning, the model "learns" a function that maps an input to an output based on labeled input-output pairs. The model makes predictions based on the patterns it has learned from the labeled data. Supervised learning incorporates classification and regression algorithms. In classification tasks, the machine learning algorithm must categorise new observations. In regression tasks, the machine learning algorithm estimates the relationship between variables. Regression focuses on a range of explanatory variables and a predicted variable, hence making it particularly relevant for forecasting. Both classification and regression play an important role in this research and the algorithms used will be discussed in sections 3.5.1 and 3.5.2.

In contrast to supervised learning, unsupervised learning does not require labeled data. Here, the algorithm interprets the data in some way to describe its structure. Common unsupervised learning tasks include some dimensionality reduction techniques and clustering.

In reinforcement learning an algorithm is given a set of actions, parameters and outcomes. The algorithm explores a variety of options to achieve the best possible outcome, adapting its approach by learning from past experiences. It is not trained on labeled input-output pairs.

3.5.1 Classification models

This section sets out the classification algorithms applied in this study.

3.5.1.1 Logistic Regression

Logistic Regression predicts the probability P that an event occurs. The algorithm employs the sigmoid function to squeeze the output of a linear equation between 0 and 1, solving

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} \tag{3.8}$$

for the coefficients β_i.

Logistic regression performs well with linearly separable data. Its coefficients provide insights into the role of independent variables as they can be interpreted as feature importances. The algorithm has low complexity and is not prone to overfitting. However, it struggles to model complex, non-linear relationships (Dreiseitl and Ohno-Machado, 2002).

3.5.1.2 Naive Bayes classifier

The Naive Bayes Classifier is based on the Bayes' theorem, which provides the posterior probability of an event given what is known as prior knowledge.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.9}$$

Naive Bayes assumes that features are independent, calculating the conditional probability as product of individual component probabilities (Bonaccorso, 2017). This assumption rarely holds true in reality.

3.5.1.3 Support Vector classifier

The Support Vector classifier ingests separable variables and builds a hyperplane that best distinguishes the two classes, creating a decision boundary that maximises the margins for both classes. To achieve this, the algorithm uses the radial basis kernel function to compute the inner products between the images of all pairs of observations. This permits projecting the data onto a higher dimensional space without calculating the actual data coordinates. The algorithm solves

$$\min_{w \in \Re^d, \xi \in \Re^+} = \frac{1}{2}||w||^2 + C\sum_i^n \xi_i \tag{3.10}$$

where w are the parameters, C is a hyperparameter controlling the penalty for misclassification and ξ_i is a slack variable.

The Support Vector classifier is useful for both linear and non-linear problems and performs well with high dimensional data (Dreiseitl and Ohno-Machado, 2002). However, the algorithm struggles with tasks related to natural language processing such as sequence classification as it has difficulty accomodating word embeddings.

3.5.1.4 Random Forest classifier

The Random Forest algorithm creates an ensemble of decision trees, trained with random subsets of features and bootstrapped data. Each tree votes by predicting the target class, and the votes are tallied to reach a final decision on the outcome. This technique is referred to as “bagging”.

This algorithm performs well with non-linear data. As long as there is a sufficient number of trees in the ensemble, overfitting is not an issue (Oughali et al., 2019). The main limitation of the Random Forest algorithm is that a large number of trees can slow the algorithm down and make it ineffective for real-time predictions.

3.5.1.5 XGBoost classifier

The XGBoost algorithm adds decision trees (also referred to as “weak learners”) sequentially to the ensemble to correct prediction errors from previous models until no improvements are made. The XGBoost algorithm stands out for its execution speed and performance (Chen and Guestrin, 2016). Tree-based algorithms such as RandomForest or XGBoost struggle to learn very complex data representations.

3.5.1.6 Multilayer Perceptron classifier

A multilayer perceptron is a simple version of a neural network in which data and calculations flow in a single direction through a number of single perceptron (also referred to as neuron) layers, from input to output layer. Neurons are the computational units that make up the “building blocks“ for neural networks, having weighted inputs and generating an output using an activation function. An activation function maps the summed weighted input to the output of the neuron. In classification tasks, the sigmoid function is used to arrive at output values between 0 and 1. First, input data is processed through all layers, computing output values and the error between the output and the expected values. The error is then propagated back through each layer of the network, updating the weights according to their error contribution. This algorithm optimises the log-loss function using stochastic gradient descent, which is a first order iterative optimisation method to find the minimum of the log loss function that measures the difference between predicted and expected output. Log-loss measures the performance of a binary classifier that predicts probabilities between 0 and 1. The more the predicted probabilities differ from the actual values, the higher the log-loss. To identify the local minimum, steps proportional to the negative of the gradient of the function at its current point are taken.

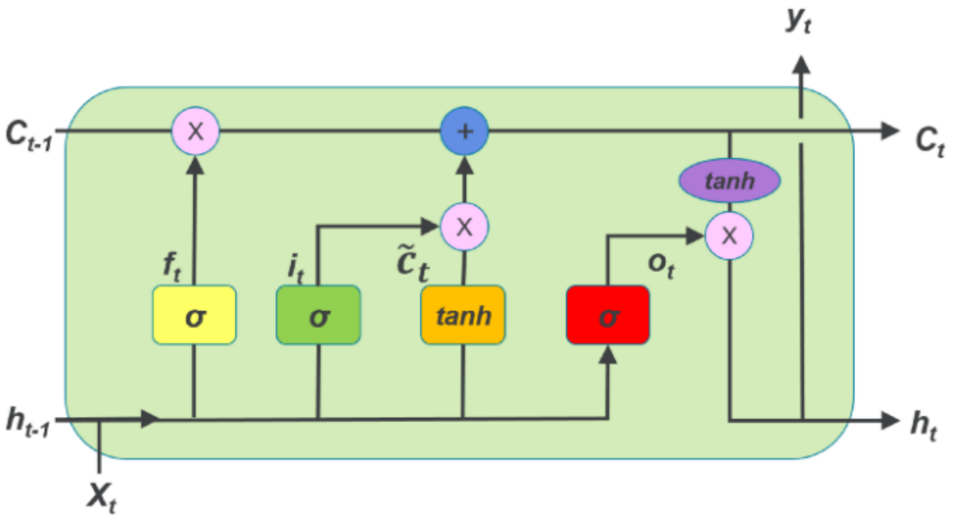
A key strength of the multilayer perceptron algorithm is its ability to learn complex representations of data (Dreiseitl and Ohno-Machado, 2002). The multilayer perceptron algorithm works well for data points that are independent from each other but struggles dealing with sequential data.

3.5.1.7 Recurrent Neural Networks

A recurrent neural network (RNN) is a special type of neural network designed for sequential data, addressing a shortcoming of basic feed-forward neural networks such as described in 3.5.1.6. RNNs have the ability to store the states or information of previous inputs to generate the next output of the sequence. A drawback of RNNs is the issue of vanishing gradients, which occurs when the gradients become very small during backpropagation, resulting in very small weight updates (see section 3.5.1.6 for details on gradient descent and backpropagation). This means that the neural network is no longer learning.

The long short term memory (LSTM) is a neural network architecture that was first proposed by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber, 1997). This type of recurrent neural network has the ability to retain information over a longer period of time while maintaining its short term performance. The LSTM architecture addresses the vanishing gradient problem by enforcing constant error flow during backpropagation through internal states of special LSTM units depicted in Fig. 3.1.

Figure 3.1: LSTM structure



The input (i_t), output (o_t) and forget (f_t) gates decide when information is added to memory, when it is output, and when it is forgotten. Through this structure, the LSTM is able to learn longer-term dependencies. Source: Ismail et al. (2018)

The output of the LSTM unit is given by

$$i_t = \delta(w_i[h_{t-1}, x_t] + b_i \tag{3.11}$$

$$f_t = \delta(w_f[h_{t-1}, x_t] + b_f] \tag{3.12}$$

$$o_t = \delta(w_o[h_{t-1}, x_t] + b_o] \tag{3.13}$$

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{3.14}$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \tag{3.15}$$

$$h_t = o_t \tanh(c_t) \tag{3.16}$$

where δ is a sigmoid function, w_x is the weight for the respective gate neurons, h_{t-1} is the output from the previous LSTM block $t-1$, x_t is the input at the current timestamp t, b_x represents the biases for the respective gates, c_t is the cell state at the current timestamp t, $\tilde{c}_t$ is the candidate for the cell state at current timestamp t and h_t is the final output.

Unidirectional LSTM models keep information from one direction as data is run forward only. A bi-directional LSTM (Bi-LSTM) propagates input data in two directions, forwards and backwards. Combining the two hidden states allows a Bi-LSTM to retain information from both ways at any point (Schuster and Paliwal, 1997). According to Graves and Schmidhuber, the Bi-LSTM architecture performs well in problems where context matters, such as sequence classification (Graves and Schmidhuber, 2005).

The Bi-LSTM algorithm is the preferred algorithm for sequence classification in this research due to its ability to learn longer-term dependencies (see chapter 4).

3.5.1.8 Hyperparameter tuning

Model parameters refer to a model's configuration that are learnt from the training data. Examples include coefficients in logistic regression or the weights in a neural network. Hyperparameters refer to external configuration of the model that are set manually, such as the learning rate in a neural network or the number of trees in random forest. Techniques such as Gridsearch and Randomsearch involve finding the best combination of hyperparameters that optimises a model's performance.

Gridsearch uses an array of values for each hyperparameter, trains and evaluates models using all possible combinations of given hyperparameter values. In Randomsearch, a statistical distribution for each hyperparameter is defined, from which values are randomly selected. Here, the number of iterations has to be determined considering time and computational resources.

Randomsearch outperforms Gridsearch on very large data sets, in which scenario Gridsearch becomes time-consuming and computationally expensive (Bergstra and Bengio, 2012). However, Gridsearch delivers equivalent results with small data sets and sufficient resources. In this research, the data sets modeled are relatively small comprising thousands of observations. Therefore, hyperparameters are optimised using Gridsearch.

3.5.2 Vector autoregression

Autoregression uses linear regression on past values to forecast the next value in a time series (see Eq. (3.3)). Vector Autoregression is a multivariate forecasting algorithm that is appropriate in scenarios where two or more time series influence each other. This algorithm models each variable as a linear combination of past values of itself and the past values of other variables in the system (Shiller and Beltratti, 1992).

Vector Autoregression differs from other autoregressive models in that the latter are unidirectional with explanatory variables influencing the predicted variable and not vice-versa. In contrast, Vector Autoregression is bi-directional in the sense that the variables in the system influence each other. In this thesis, Granger Causality is used to test for the bi-directional nature of relationships between variables (see section 3.2 for a detailed discussion).

A vector autoregressive framework permits modeling a $T \times K$ multivariate time series Y, where T stands for the number of observations and K the number of variables. The framework is defined as

$$Y_t = v + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + u_t \tag{3.17}$$

where A_i denotes a $K \times K$ coefficient matrix, v a constant and u_t white noise.

This research attempts to forecast economic indicators using variables that exhibit bi-directional relationships and thus employs a vector autoregressive model for this purpose (see chapters 4 and 5).

3.5.2.1 Model selection

While at times, economic theory can inform the number of lags in autoregressive frameworks, this research relies on statistical methods to determine the appropriate number of lags that should be included as regressors. This is important as too many lags inflate the standard errors of coefficient estimates and imply an increase in the forecast error while too few lags can result in an estimation bias.

The Akaike (AIC), Bayesian (BIC) and the Hannan-Quinn (HQIC) information criteria are methods for obtaining the optimal lag length. These metrics are based on the concept that adding a further term may improve the model while adding a penalty for increasing the number of parameters. When the increase in goodness-of-fit is larger than the penalty term, the information criterion statistic decreases. Therefore, the model with the lowest scores should be selected (Brooks and Tsolacos, 2010). AIC is defined as

$$AIC = -2\frac{\ell}{n} + 2\frac{k}{n} \tag{3.18}$$

where n represents the number of observations, k denotes the number of estimated model parameters and ℓ stands for the log likelihood function. The log likelihood function is given by

$$\ell = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)\right) \tag{3.19}$$

BIC is given by

$$BIC = -2\frac{\ell}{n} + \frac{k \times \ln n}{n} \tag{3.20}$$

where n denotes the number of observations, k stands for the number of estimated model parameters and ℓ stands for the log likelihood function.

The Hannan-Quinn Information criterion (HQIC) is defined as

$$HQIC = -2\frac{\ell}{n} + 2 \times \frac{k \times \ln(\ln n)}{n} \tag{3.21}$$

where n denotes the number of observations, k stands for the number of estimated model parameters and ℓ stands for the log likelihood function. Glaeskens and Hjort show that $\ln n$ is not the slowest rate by which the penalty can increase to infinity to almost certainly choose the most parsimonious model. In that respect, HQIC applies the law of the iterated logarithm to ensure strong consistency in model selection (Claeskens et al., 2008).

AIC has a lower penalty than BIC, causing AIC to select more complex models. Compared to AIC, BIC has a higher penalty for model complexity, resulting in more complex models being less likely to be selected. The probability for BIC and HQIC selecting the "true model" increases with the size of the data set given that the scores are based on the Bayesian probability concept, which is not the case for AIC. For small data sets, BIC and HQIC are more likely to choose models that are too simple. According to Yang, AIC is asymptotically efficient for choosing the model with the smallest mean squared error in regression tasks, assuming that the "true model" does not exist. BIC and HQIC are not asymptotically efficient under this assumption (Yang, 2005).

The Minimum Description Length (MDL) is a further method for model selection, approaching the task from an information theory perspective. Information theory is concerned with the representation and transfer of data. The MDL score stands for the smallest number of bits needed to describe the data and the model (Witten et al., 2005). Like MDL, BIC can be viewed as means to select a model by minimum description length. A drawback of MDL is its computational expensiveness.

Forecasters use one or more model selection methods, with AIC and BIC the most commonly used approaches (Kapetanios and Papailias, 2018). In this research, model selection is based on AIC and BIC scores, with more weight given to the latter score to avoid overfitting in case of disagreement (see chapters 4 and 5).

3.6 Assessing model performance

This section reviews methods for evaluating model performance.

3.6.1 Classification performance metrics

When selecting metrics for evaluating a classification model, it is important to reflect on what problem the model is attempting to solve. In this research, it is important that the classifiers perform well in identifying positive examples.

Precision measures a classifier's ability to correctly identify positive examples.

$$Precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ and\ false\ positives} \tag{3.22}$$

Recall measures a classifier's ability to find positive examples.

$$Recall = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ and\ false\ negatives} \tag{3.23}$$

The F1 metric is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.24}$$

The metrics set out in Eqs. (3.22), (3.23) and (3.24) take values between zero and one, with the latter describing the perfect classifier.

An alternative to the above metrics is accuracy, which is the ratio of correct predictions and total predictions. However, accuracy is an inappropriate metric when classes are imbalanced. For instance, if there is a low ratio of positive observations in the data set, the metric will not convey much about the classifier's power to correctly predict the minority class.

He and Ma argue that precision, recall and F1 metrics are suitable for information retrieval tasks as they reveal the portion of relevant data retrieved along with the quantity data correctly identified as relevant (He and Ma, 2013). Indeed, they are more appropriate metrics for the evaluation of a classifier than accuracy, in particular when the classes are imbalanced, avoiding a bias towards the dominant class. Precision, recall and F1 metrics are used in chapters 4, 5 and 6.

3.6.2 Regression performance metrics

A model's goodness of fit can be evaluated with the root mean squared error (RMSE) metric, which calculates the difference between predicted and actual values, estimating the standard deviations of the error distribution (Hoffmann et al., 2019). For *n* forecasts in a time series, RMSE is computed as the square root of the mean of the residuals' squares.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \tag{3.25}$$

where n stands for the number of observations and x_i and $\hat{x}_i$ are the actual and predicted values, respectively.

An alternative metric to assess the goodness of fit in regression models is the coefficient of determination or R^2. This metric denotes the proportion of the variation in the predicted variable that is explained by the explanatory variables and takes values between 0 and 1.

When evaluating time series forecasting models, R^2 assesses how well a model fits past values, however no conclusions can be made regarding model performance on unseen, or future, data. The coefficient of determination does not inform whether the selected model is appropriate or whether predictions are biased. Therefore, in this research, the preferred metric for evaluating time series forecasts is the RMSE (see chapters 4 and 5).

3.6.3 Cross-validation

Cross-validation is a method that shows how well a model generalises on unseen data and describes the iterative process of resampling the data for training and testing.

With the holdout method, a data set is split into train and test sets. However, model performance is very sensitive to the manner in which the data is split and can therefore vary notably for the two sets.

K-fold cross-validation addresses this drawback by splitting the training data set randomly into k subsets and rotating training data on all subsets apart from one that is held out, and assessing model performance on the held out validation data. The procedure is performed until all subsets have been the held out validation set, thus providing a robust estimate of model performance. For each fold, model performance E_1, E_2, ..., E_k is calculated, with the overall performance defined as $E = \frac{1}{k}\sum_{i=1}^{k} E_i$.

When modelling time-series data, the data set cannot be split into k random chunks given the temporal dependency between observations. Instead, performance is evaluated using walk-forward cross-validation. For walk-forward cross-validation, the data set is divided into k parts. In the k^{th} split, the first k folds are returned as train set and the $(k+1)^{th}$ fold is used as test set. By design, this method assumes that past data is relevant.

In this thesis, k-fold cross-validation and walk-forward cross-validation in the case of time-series data are employed to generate robust model performance estimates (see chapters 4, 5 and 6).

3.6.4 Comparing model predictions

This section reviews tests for multiple time series forecast comparison as well as methods for comparing classifier predictions.

The tests for time series forecast comparison considered are model-free, i.e. the model that generated the forecasts does not have to be available. They all test the null hypothesis that the forecasts are not significantly different.

The Morgan-Granger-Newbold test stipulates that forecasts are not significantly different if the expectation for the loss differential is zero for all observations t in a time series. The test assumes that the loss is quadratic and that the errors have a zero mean, are Gaussian and serially uncorrelated. Let e_i and r_i be the residuals of two forecasts.

$$x_t = e_i + r_i \tag{3.26}$$

$$z_t = e_i - r_i \tag{3.27}$$

The null hypothesis is not rejected if the two forecast error variances set out above are equal or the covariance between x_t and z_t is zero.

The Meese Rogoff test allows for serial and contemporaneous correlation in forecast errors. Like the Morgan-Granger-Newbold test, it assumes the loss is quadratic and that the errors are Gaussian with a zero mean. The Meese Rogoff test is based on the sample covariance of x_t and z_t.

The key drawback of both tests consists in the handling of non-Gaussian forecast errors, which are significantly missized in both large and small data sets. The following method, namely the Diebold Mariano test, has the ability to retain the correct forecast error size with the exception of small data samples (Mariano, 2002).

The Diebold Mariano test assumes that the time series loss differential d_i is stationary (Diebold and Mariano, 2002). The loss differential is given by

$$d_i = e_i^2 - r_i^2 \tag{3.28}$$

The sample mean loss differential $\bar{d_i}$ is then defined as

$$\bar{d_i} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{3.29}$$

where n refers to the sample size. The spectral density of the loss differential at frequency 0 is given by

$$f_d(0) = \frac{1}{2\pi}\left(\sum_{k=-\infty}^{\infty} \gamma_d(k)\right) \tag{3.30}$$

where $\gamma_d(k)$ stands for the autocovariance of the loss differential at lag k. The Diebold Mariano test statistic is given by

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi\hat{f}_d(0)}{n}}} \tag{3.31}$$

where $\hat{f}_d(0)$ is a consistent estimator of $f_d(0)$. The difference between two forecasts is statistically significant if $|DM| > zcrit$ where $zcrit$ is the two-tailed critical value for the standard normal distribution. However, a normal distribution can be a poor proxy for small samples where the Diebold Mariano test tends to be too conservative.

Harvey, Leybourne and Newbold (Harvey et al., 1997) address this shortcoming by adjusting the Diebold-Mariano statistic, making it more appropriate for smaller sample sizes

$$HLN = DM\sqrt{\frac{n+1-2h+h(h-1)}{n}} \tag{3.32}$$

where n denotes the sample size and h stands for the steps-ahead-forecast. The HLN statistic is compared with a student-t distribution with $(n-1)$ degrees of freedom instead of a standard normal distribution.

In this research, the adjusted Diebold Mariano test is the preferred method for comparing multiple regression model forecasts (see chapters 4 and 5).

A method for comparing the predictions of two classifiers is the McNemar test (Dietterich, 1998). This test is based on a contingency table, whose structure in case of a binary classification is set out in Table 3.1.

Table 3.1: Contingency table

Outcome	Classifier 2 correct	Classifier 2 incorrect
Classifier 1 correct	a	b
Classifier 1 incorrect	c	d

The cell entries of the contingency table stand for the count of observations in each of the four cases. Both classifiers are trained on exactly the same training and test sets, respectively.

The McNemar test evaluates the disagreements between two models' predictions. The McNemar statistic is defined as

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{3.33}$$

χ^2 has a chi-squared distribution with one degree of freedom. If the statistic is larger than a set α level, then the null hypothesis cannot be rejected and the two classifier predictions are considered not significantly different.

The McNemar test requires only one fit for each of the classifiers, which is is advantageous when working with large data sets and models that are slow to train. The test makes no assumption about the distribution of data.

Two alternatives to the McNemar test include the Two Proportions test and the Paired T-test. The Two Proportions test is applied to the data of two random samples, each independently drawn from a different population. This test assumes the independence of samples, which is not the case when comparing the predictions of two classifiers as the train and test sets are the same for both classifiers. The Paired T-test evaluates if the mean of a predicted variable is the same for two model predictions. The test assumes normally distributed data.

The McNemar test has the lowest type I error among tests that only need to be run once and does not make assumption of the underlying data distribution (Dietterich, 1998). It is therefore the preferred test for comparing classifier predictions in this research (see chapter 6).

3.7 Graph principles and metrics

This section outlines the graph concepts that are applied in this research.

Graphs are objects that represent networks of distinct elements, i.e. nodes and a set of edges or links that connects some or all node pairs. In an undirected graph, edges do not represent directionality, while edges in a directed graph are distinct ordered pairs of nodes. The node degree describes the sum of edges incident in node k_i. Directed graphs have in- and outdegrees and in such case, the node degree is the sum of $k_{in} + k_{out}$.

Two nodes are connected if a path exists between them. An undirected graph is connected if every pair of nodes is connected. A directed graph is weakly connected if substituting all of its directed edges with undirected edges results in a connected, undirected graph. A directed graph is strongly connected if it includes a directed edge from u to v and a directed path from v to u for every pair of nodes u, v. Weighted graphs have a numerical value associated with each edge. In such graphs, the node strength is the sum of weights of links incident in node i.

3.7.1 Graph centrality metrics

An adjacency matrix is one way to represent a graph. In an unweighted graph, this is an $N \times N$ matrix with entries a_{ij}. Entries are zero or one depending if a link exists between nodes i and j. In case of weighted graphs, ones are replaced by the numerical value of the edge weight.

In an undirected graph, the degree centrality is defined as the number of edges incident in a node. The concept of degree centrality can be extended to measuring a node's influence in a graph using the eigenvector centrality metric. A node with a high eigenvector centrality is linked to other high-scoring nodes which impact this metric more than an equal amount of low-scoring nodes. Hence, this measure describes the overall importance of a node in the graph and is suitable for connected graphs. The eigenvector centrality for node c_i^E in an undirected, connected graph is defined as

$$c_i^E = u_{1,i} \tag{3.34}$$

where $u_{1,i}$ stands for the i^{th} component of u_1, the normalised eigenvector associated with the leading eigenvalue λ_1 of adjacency matrix A. Based on the Perron–Frobenius theorem, only the largest non-negative eigenvalue results in the centrality measure, thus ensuring that the definition of centrality has meaning.

A further centrality measure, namely the betweenness centrality, is based on the idea that a node is central if it is located between many other nodes.

This research utilises eigenvector centrality as preferred node centrality metric due to its emphasis on the overall importance of a node in the graph structure (see chapter 5).

3.7.2 Graph similarity metrics

There are various methods for comparing graphs. The Jaccard Similarity Coefficient compares members of two sets (e.g. of graph nodes) to assess which members are shared and which ones are not. One drawback of the Jaccard Coefficient is that it is not sensitive to set sizes. The Overlap Coefficient attempts to address this shortcoming and is defined as the size of the union of set A and set B over the size of the smaller set between A and B.

The Portrait Divergence metric uses a graph's portrait based on the distribution of the shortest-path lengths in a graph (Bagrow and Bollt, 2019). A graph portrait is a matrix B whose item $B_{l,k}$, l = 0, 1, ..., d (d = graph diameter), k = 0, 1, $N-1$ (i.e. the amount of nodes with k nodes of shortest path distance l). This idea is also relevant to weighted graphs where a binning strategy is applied to account for path lengths. The portrait divergence score is calculated in two stages. First, the probability $P(k,l)$ of choosing two nodes at distance l at random and, for one of the two nodes, to have k nodes at distance l, is calculated for the two graphs to be compared:

$$P(k,l) = P(k|l)P(l) = \frac{1}{N}B_{lk}\frac{1}{\sum_c n_c^2}\sum_{k'=0}^{N} k'B_{lk'} \tag{3.35}$$

where n_c is the number of nodes in the connected structure c. Then, the distance between the two graphs' portraits is calculated as follows:

$$D(G_1,G_2) = \frac{1}{2}KL(P||M) + \frac{1}{2}(KL(Q||M) \tag{3.36}$$

where $M = (P+Q)/2$ is the mixture distribution of P and Q (as set out in Eq. (3.35)), and $KL(\cdot||\cdot)$ represents the Kullback-Leibler divergence.

The portrait divergence metric represents a comprehensive metric of how the topological properties of two graphs differ (Tantardini et al., 2019) and is therefore the preferred graph similarity measure in this research (see chapter 5).

3.7.3 Graph filtering methods

Knowledge graphs can at times be very large and noisy, which requires the identification of statistically relevant pairs of edges, referred to as the “backbone”. Existing approaches for extracting graph backbones can be grouped into “coarse-grained” and filter-based methods (Ghalmane et al., 2020). “Coarse-grained” approaches build on the idea of classifying graph nodes according to some criterion and retaining those with the desired attributes, while filter-based approaches set out attributes for nodes and edges and reject or retain them based on the statistical significance of these properties against a null hypothesis for the graph structure. The main drawback of the “coarse-grained“ methods is that they may result in an unintended loss of information.

In terms of the filter-based methods, the hypergeometric filter was introduced for bipartite graphs and retains edges that are statistically significant with regard to a null hypothesis of random connections in a weighted graph, characterised by a hypergeometric distribution (Tumminello et al., 2011). Another filtering technique for weighted graphs is the disparity filter (Serrano et al., 2009). This filter normalizes each edge weight (i,j) considering the strength of node i (which stands for the sum of weights of edges connected to node i) to add up to one. The filter’s null hypothesis specifies that the normalised weights associated to the edges of a node i are random and uniformly distributed over $[0,1]$. Then, a p-value is computed, representing the probability α_{ij} of the edge complying with the null hypothesis. The p-value is defined as

$$\alpha_{ij} = (1 - p_{ij})^{k-1} \tag{3.37}$$

where p_{ij} stands for the normalised weight and k represents the node degree.

The null hypothesis is rejected for those edges whose p-value is below a set threshold α and they are preserved in the backbone.

Both hypergeometric and disparity filters have been empirically explored and the disparity filter is the preferred method for backbone extraction in this research as it generates more parsimonious backbones (see chapter 5).

Chapter 4

Macroeconomic forecasting through news, emotions and narrative

4.1 Introduction, motivation and background

This study contributes to existing literature by including a wide range of emotions from global newspapers into macroeconomic forecasts using data from the Global Database of Events, Language and Tone (GDELT) (Leetaru, 2015b). I propose a filtering methodology based on machine learning to identify articles pertinent to economic growth and inflation, and show that emotions expressed in those news articles improve forecasts of industrial production and consumer prices across a diverse range of economies. I employ dimensionality reduction to summarise more than 600 emotion scores into a smaller number of interpretable components. Results suggest that emotions linked to “happiness” and “anger” have the strongest predictive power across the indicators I forecast. To the best of my knowledge, emotion scores from GDELT’s Content Analysis Measure Systems have not yet been used to predict macroeconomic indicators.

4.2 Literature review

This section discusses a choice of existing research relevant to macroeconomic forecasting with news sentiment.

A rapidly growing body of literature explores the use of media sentiment and big data for economic predictions (Buono et al., 2017; Kapetanios and Papailias, 2018). Most works predict macroeconomic indicators with regression frameworks including traditional data alongside positive and negative tone based on word count (in contrast to a wider array of emotions).

Established works find that newspaper sentiment is an effective means for tracking changes in the economic cycle (Tuckett et al., 2014; Shiller, 2019). Likewise, unsettling news reports herald change in the economy, suggesting that news could be utilised for risk management (Nyman et al., 2021).

Baker *et al.* have created indices of policy uncertainty for a broad range of geographies (Baker et al., 2016, 2020). The authors leverage an autoregressive framework incorporating news-based features and macroeconomic variables to determine whether uncertainty shocks are predictive of an economic slowdown. Results indicate that policy uncertainty increases stock market volatility, decreases investment returns and employment growth. Political bias does not significantly affect these indices. Nyman and Ormerod (Nyman and Ormerod, 2020) derive uncertainty-related terms from Reuters news via natural language processing techniques and demonstrate that they have a causal link with the uncertainty index developed by Baker *et al.* (Baker et al., 2016).

Fraiberger *et al.* use Reuters news to derive positive and negative tone scores (Fraiberger et al., 2018). The authors explain that news sentiment improves forecasts for developed and emerging equity markets. Global news sentiment has a larger effect on global stock markets than local sentiment.

Thorsrud shows that including news topics from raw news narrative into forecasts of quarterly GDP changes leads to significant improvements compared to central bank predictions (Thorsrud, 2016). Larson and Thorsrud build on previous work demonstrating that newspaper topics have strong predictive power for Norwegian economic indicators and stock market prices (Larsen and Thorsrud, 2019).

A paper by Pekar and Binner shows that incorporating data on planned purchases from Twitter tweets alongside lagged consumer index values produces statistically significant improvements of consumer price forecasts over the baseline model that solely includes lag variables (Pekar and Binner, 2017).

Fronzetti Colladon *et al.* create a sentiment index based on the significance of economic keywords in Italian newspapers and demonstrate that this index is able to forecast Italian equity and fixed income market volatilities and returns, including during the COVID-19 pandemic in 2020 (Fronzetti Colladon et al., 2020).

Newspaper archives and Twitter are frequently utilised sources for raw textual data. Nonetheless, there is a growing body of literature leveraging preprocessed sentiment scores. Ortiz blends official economic statistics with themes from GDELT to monitor Chinese economic weakness in real-time, providing a better awareness of changes in the Chinese economy for policymakers and investors (Casanova et al., 2017). Elshendy *et al.* combine data from GDELT together with conventional macroeconomic indicators and use social network analysis to forecast macroeconomic indices such as consumer confidence, business confidence and GDP for 10 large EU countries (Elshendy and Fronzetti Colladon, 2017). Findings suggest that information from GDELT adds value to predictions of macroeconomic variables.

Chen explores the impact of the negative news coverage with regard to international trade from US presidential candidates in 2016 using the average tone score from GDELT (Chen and Lo, 2019). The paper finds that news narrative has the power to impact the economy by affecting market participants' expectations.

Glaeser *et al.* incorporate YELP reviews into forecast of the local economy and demonstrate that this information adds value to predictions of coincidental changes in the local economy (Glaeser et al., 2017). The authors concede that, due to the large amount of reviews, using YELP data yields the best results when applied to urban areas and the hospitality industry.

Most works find that incorporating news sentiment into macroeconomic forecasting frameworks adds value. However, Schaer *et al.* highlight the importance of sound statistical testing and deliberate selection of error metrics and benchmarks and note some of the issues when working with sentiment data such as data complexity, sampling instability and choice of key words (Schaer et al., 2019).

The majority of research focuses on positive and negative sentiment for improving macroeconomic predictions, mainly for the US, with only a few papers incorporating a wider array of emotions in their analyses.
In this chapter, I attempt to forecast industrial production (IP) and consumer prices (CPI) for 10 diverse economies using a wide range of nuanced emotion scores from GDELT. I address one research question, i.e., whether emotion scores from GDELT add value to forecasts of IP and CPI. Accordingly, I formulate the following hypothesis:

- H_1: Changes in GDELT emotion scores are predictive of changes in IP and CPI.

This experiment expands the existing body of literature by incorporating nuanced sentiment scores extracted from global newspaper narrative into macroeconomic forecasts of industrial production and consumer prices for 10 diverse economies. I complement the results with an interpretability study, setting out which emotions have the strongest predictive power.

4.3 Data and methods

This section describes the data source used, the predicted variables, the filtering methodology and the data sets created. Further, it sets out the nature of the sentiment scores used in this experiment and outlines the preprocessing steps.

4.3.1 The Global Database of Events, Language and Tone (GDELT)

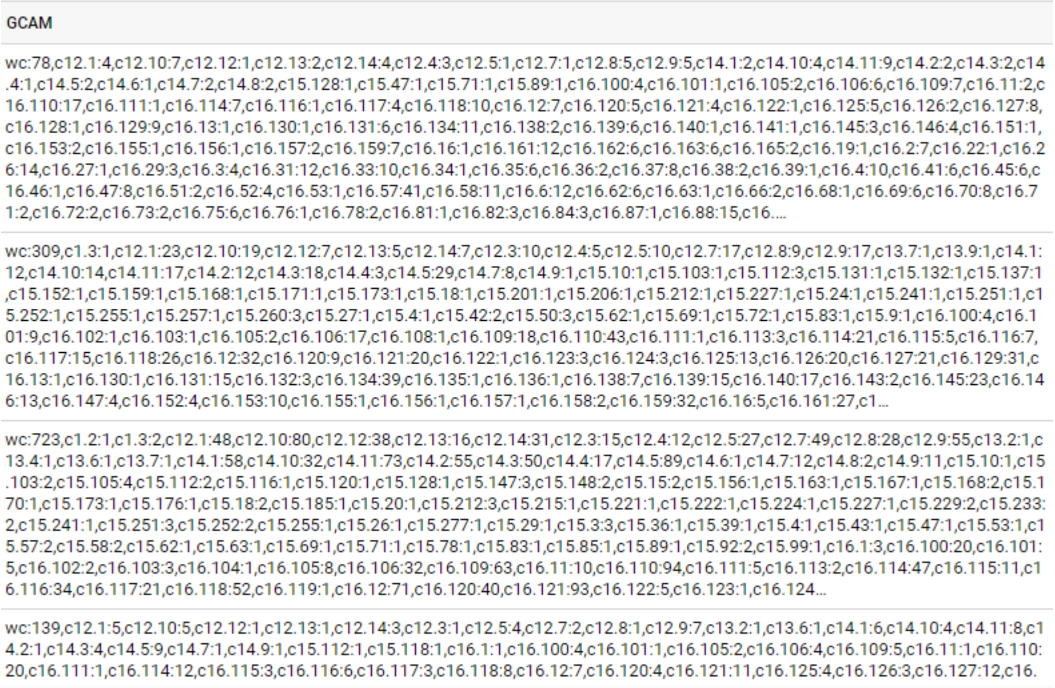
This research uses GDELT as source of narrative-based features.

The GDELT Project is a joint research undertaking of different bodies at Google, the Yahoo! Fellowship at Georgetown University and several large news archives such as BBC Monitoring or LexisNexis. The project monitors world media from a multitude of perspectives, identifying and extracting items such as themes, emotions, locations and events. GDELT version two incorporates real-time translation from 65 languages and measures over 2,300 emotions and themes from every news article, updated every 15 minutes (Leetaru, 2015b). It is a public data set available on the Google Cloud Platform.

The Global Knowledge Graph (GKG), one of the tables within GDELT, includes fields such as sentiment scores, themes and locations extracted from global newspaper articles. This table is populated by a software that analyses global newspaper articles in real-time to identify entities such as themes, emotions, locations, organizations, and many more (Leetaru et al., 2014).

Within GDELT's GKG, the Global Content Analysis Measures (GCAM) field includes over 2,300 emotion scores. The GCAM system includes 24 content analysis systems that scan each news item and feed the computed emotion scores as a comma-delimited string into the GCAM field. Most of the GCAM scores are word count-based, with a few generated by more sophisticated approaches. Besides the emotion scores, the GCAM field contains an overall word count for every article scanned. Fig. 4.1 shows a sample of the data in the GCAM field.

Figure 4.1: Sample of GCAM field

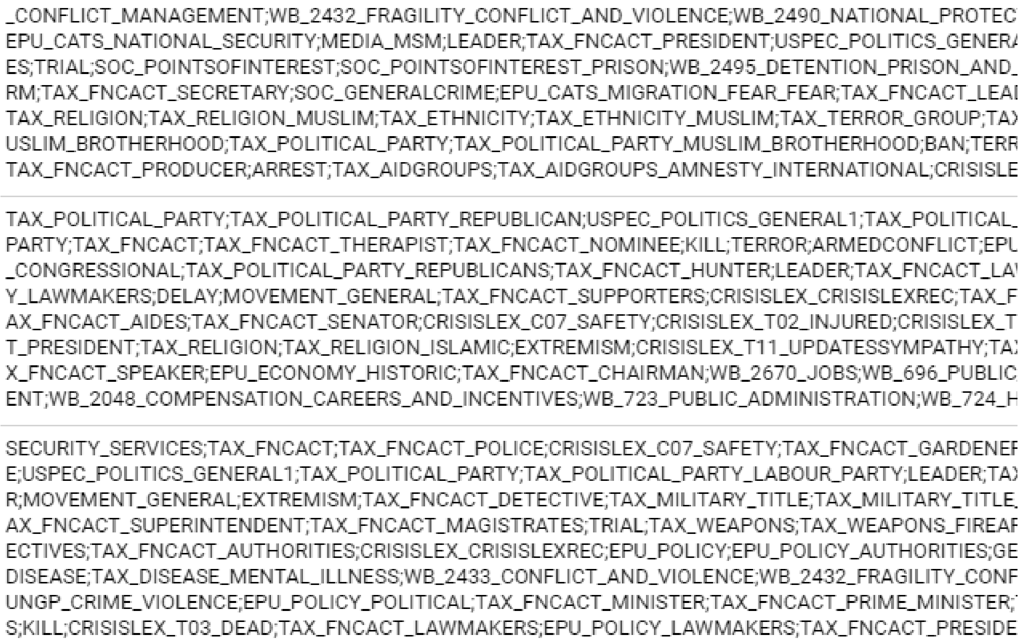


GCAM
wc:78,c12.1:4,c12.10:7,c12.12:1,c12.13:2,c12.14:4,c12.4:3,c12.5:1,c12.7:1,c12.8:5,c12.9:5,c14.1:2,c14.10:4,c14.11:9,c14.2:2,c14.3:2,c14.4:1,c14.5:2,c14.6:1,c14.7:2,c14.8:2,c15.128:1,c15.47:1,c15.71:1,c15.89:1,c16.100:4,c16.101:1,c16.105:2,c16.106:6,c16.109:7,c16.11:2,c16.110:17,c16.111:1,c16.114:7,c16.116:1,c16.117:4,c16.118:10,c16.12:7,c16.120:5,c16.121:4,c16.122:1,c16.125:5,c16.126:2,c16.127:8,c16.128:1,c16.129:9,c16.13:1,c16.130:1,c16.131:6,c16.134:11,c16.138:2,c16.139:6,c16.140:1,c16.141:1,c16.145:3,c16.146:4,c16.151:1,c16.153:2,c16.155:1,c16.156:1,c16.157:2,c16.159:7,c16.16:1,c16.161:12,c16.162:6,c16.163:6,c16.165:2,c16.19:1,c16.2:7,c16.22:1,c16.26:14,c16.27:1,c16.29:3,c16.3:4,c16.31:12,c16.33:10,c16.34:1,c16.35:6,c16.36:2,c16.37:8,c16.38:2,c16.39:1,c16.4:10,c16.41:6,c16.45:6,c16.46:1,c16.47:8,c16.51:2,c16.52:4,c16.53:1,c16.57:41,c16.58:11,c16.6:12,c16.62:6,c16.63:1,c16.66:2,c16.68:1,c16.69:6,c16.70:8,c16.71:2,c16.72:2,c16.73:2,c16.75:6,c16.76:1,c16.78:2,c16.81:1,c16.82:3,c16.84:3,c16.87:1,c16.88:15,c16....
wc:309,c1.3:1,c12.1:23,c12.10:19,c12.12:7,c12.13:5,c12.14:7,c12.3:10,c12.4:5,c12.5:10,c12.7:17,c12.8:9,c12.9:17,c13.7:1,c13.9:1,c14.1:12,c14.10:14,c14.11:17,c14.2:12,c14.3:18,c14.4:3,c14.5:29,c14.7:8,c14.9:1,c15.10:1,c15.103:1,c15.112:3,c15.131:1,c15.132:1,c15.137:1,c15.152:1,c15.159:1,c15.168:1,c15.171:1,c15.173:1,c15.18:1,c15.201:1,c15.206:1,c15.212:1,c15.227:1,c15.24:1,c15.241:1,c15.251:1,c15.252:1,c15.255:1,c15.257:1,c15.260:3,c15.27:1,c15.4:1,c15.42:2,c15.50:3,c15.62:1,c15.69:1,c15.72:1,c15.83:1,c15.9:1,c16.100:4,c16.101:9,c16.102:1,c16.103:1,c16.105:2,c16.106:17,c16.108:1,c16.109:18,c16.110:43,c16.111:1,c16.113:3,c16.114:21,c16.115:5,c16.116:7,c16.117:15,c16.118:26,c16.12:32,c16.120:9,c16.121:20,c16.122:1,c16.123:3,c16.124:3,c16.125:13,c16.126:20,c16.127:21,c16.129:31,c16.13:1,c16.130:1,c16.131:15,c16.132:3,c16.134:39,c16.135:1,c16.136:1,c16.138:7,c16.139:15,c16.140:17,c16.143:2,c16.145:23,c16.146:13,c16.147:4,c16.152:4,c16.153:10,c16.155:1,c16.156:1,c16.157:1,c16.158:2,c16.159:32,c16.16:5,c16.161:27,c1...
wc:723,c1.2:1,c1.3:2,c12.1:48,c12.10:80,c12.12:38,c12.13:16,c12.14:31,c12.3:15,c12.4:12,c12.5:27,c12.7:49,c12.8:28,c12.9:55,c13.2:1,c13.4:1,c13.6:1,c13.7:1,c14.1:58,c14.10:32,c14.11:73,c14.2:55,c14.3:50,c14.4:17,c14.5:89,c14.6:1,c14.7:12,c14.8:2,c14.9:11,c15.10:1,c15.103:2,c15.105:4,c15.112:2,c15.116:1,c15.120:1,c15.128:1,c15.147:3,c15.148:2,c15.15:2,c15.156:1,c15.163:1,c15.167:1,c15.168:2,c15.170:1,c15.173:1,c15.176:1,c15.18:2,c15.185:1,c15.20:1,c15.212:3,c15.215:1,c15.221:1,c15.222:1,c15.224:1,c15.227:1,c15.229:2,c15.233:2,c15.241:1,c15.251:3,c15.252:2,c15.255:1,c15.26:1,c15.277:1,c15.29:1,c15.3:3,c15.36:1,c15.39:1,c15.4:1,c15.43:1,c15.47:1,c15.53:1,c15.57:2,c15.58:2,c15.62:1,c15.63:1,c15.69:1,c15.71:1,c15.78:1,c15.83:1,c15.85:1,c15.89:1,c15.92:2,c15.99:1,c16.1:3,c16.100:20,c16.101:5,c16.102:2,c16.103:3,c16.104:1,c16.105:8,c16.106:32,c16.109:63,c16.11:10,c16.110:94,c16.111:5,c16.113:2,c16.114:47,c16.115:11,c16.116:34,c16.117:21,c16.118:52,c16.119:1,c16.12:71,c16.120:40,c16.121:93,c16.122:5,c16.123:1,c16.124...
wc:139,c12.1:5,c12.10:5,c12.12:1,c12.13:1,c12.14:3,c12.3:1,c12.5:4,c12.7:2,c12.8:1,c12.9:7,c13.2:1,c13.6:1,c14.1:6,c14.10:4,c14.11:8,c14.2:1,c14.3:4,c14.5:9,c14.7:1,c14.9:1,c15.112:1,c15.118:1,c16.1:1,c16.100:4,c16.101:1,c16.105:2,c16.106:4,c16.109:5,c16.11:1,c16.110:20,c16.111:1,c16.114:12,c16.115:3,c16.116:6,c16.117:3,c16.118:8,c16.12:7,c16.120:4,c16.121:11,c16.125:4,c16.126:3,c16.127:12,c16.

Every “box” corresponds to one news item scanned. The first score is always “wc”, which stands for the wordcount of a specific news item. The remainder consists of the GCAM sentiment scores and their respective value calculated by the GCAM analysis systems.

Location data is extracted via full text geocoding, an approach proposed by Leetaru (Leetaru, 2012). This method employs algorithms to scan newspaper narrative and to extract mentions of locations using large databases of places. Following the same idea, extensive topic lists are applied to identify themes in news reporting. Fig. 4.2 contains a sample of the themes field, with each “box” corresponding to one news item parsed.

Figure 4.2: Sample of themes field



```
_CONFLICT_MANAGEMENT;WB_2432_FRAGILITY_CONFLICT_AND_VIOLENCE;WB_2490_NATIONAL_PROTEC
EPU_CATS_NATIONAL_SECURITY;MEDIA_MSM;LEADER;TAX_FNCACT_PRESIDENT;USPEC_POLITICS_GENERA
ES;TRIAL;SOC_POINTSOFINTEREST;SOC_POINTSOFINTEREST_PRISON;WB_2495_DETENTION_PRISON_AND_
RM;TAX_FNCACT_SECRETARY;SOC_GENERALCRIME;EPU_CATS_MIGRATION_FEAR_FEAR;TAX_FNCACT_LEA
TAX_RELIGION;TAX_RELIGION_MUSLIM;TAX_ETHNICITY;TAX_ETHNICITY_MUSLIM;TAX_TERROR_GROUP;TA
USLIM_BROTHERHOOD;TAX_POLITICAL_PARTY;TAX_POLITICAL_PARTY_MUSLIM_BROTHERHOOD;BAN;TERR
TAX_FNCACT_PRODUCER;ARREST;TAX_AIDGROUPS;TAX_AIDGROUPS_AMNESTY_INTERNATIONAL;CRISISLE
```

```
TAX_POLITICAL_PARTY;TAX_POLITICAL_PARTY_REPUBLICAN;USPEC_POLITICS_GENERAL1;TAX_POLITICAL_
PARTY;TAX_FNCACT;TAX_FNCACT_THERAPIST;TAX_FNCACT_NOMINEE;KILL;TERROR;ARMEDCONFLICT;EPU
_CONGRESSIONAL;TAX_POLITICAL_PARTY_REPUBLICANS;TAX_FNCACT_HUNTER;LEADER;TAX_FNCACT_LA
Y_LAWMAKERS;DELAY;MOVEMENT_GENERAL;TAX_FNCACT_SUPPORTERS;CRISISLEX_CRISISLEXREC;TAX_F
AX_FNCACT_AIDES;TAX_FNCACT_SENATOR;CRISISLEX_C07_SAFETY;CRISISLEX_T02_INJURED;CRISISLEX_T
T_PRESIDENT;TAX_RELIGION;TAX_RELIGION_ISLAMIC;EXTREMISM;CRISISLEX_T11_UPDATESSYMPATHY;TA
X_FNCACT_SPEAKER;EPU_ECONOMY_HISTORIC;TAX_FNCACT_CHAIRMAN;WB_2670_JOBS;WB_696_PUBLIC
ENT;WB_2048_COMPENSATION_CAREERS_AND_INCENTIVES;WB_723_PUBLIC_ADMINISTRATION;WB_724_H
```

```
SECURITY_SERVICES;TAX_FNCACT;TAX_FNCACT_POLICE;CRISISLEX_C07_SAFETY;TAX_FNCACT_GARDENE
E;USPEC_POLITICS_GENERAL1;TAX_POLITICAL_PARTY;TAX_POLITICAL_PARTY_LABOUR_PARTY;LEADER;TA
R;MOVEMENT_GENERAL;EXTREMISM;TAX_FNCACT_DETECTIVE;TAX_MILITARY_TITLE;TAX_MILITARY_TITLE_
AX_FNCACT_SUPERINTENDENT;TAX_FNCACT_MAGISTRATES;TRIAL;TAX_WEAPONS;TAX_WEAPONS_FIREA
ECTIVES;TAX_FNCACT_AUTHORITIES;CRISISLEX_CRISISLEXREC;EPU_POLICY;EPU_POLICY_AUTHORITIES;GE
DISEASE;TAX_DISEASE_MENTAL_ILLNESS;WB_2433_CONFLICT_AND_VIOLENCE;WB_2432_FRAGILITY_CONF
UNGP_CRIME_VIOLENCE;EPU_POLICY_POLITICAL;TAX_FNCACT_MINISTER;TAX_FNCACT_PRIME_MINISTER;
S;KILL;CRISISLEX_T03_DEAD;TAX_FNCACT_LAWMAKERS;EPU_POLICY_LAWMAKERS;TAX_FNCACT_PRESIDE
```

Every “box” corresponds to one news item scanned. Themes are extracted as strings of labels identified by the GDELT algorithm when scanning a news item.

The theme field includes around 13,000 unique, very nuanced themes, which can be grouped into higher-level categories such as “health”, “government” or “environment” using the themes’ taxonomy.

The GKG accounts for around 12 terabytes of data drawn from c 47,000 global news sources, with new information being added constantly, starting end of February 2015. GKG focuses on newspapers with an online presence and does not include data from social media platforms. To date, GDELT has analysed in excess of one billion news articles.

4.3.2 Predicted variables

This experiment forecasts industrial production (IP) and consumer price indices (CPI) for US, UK, Germany, Norway, Poland, Turkey, Japan, South Korea, Brazil and Mexico.

IP is a monthly proxy of economic activity. It captures the output of industrial enterprises and monitors the change in production output quantity.

The consumer price index (CPI) is chosen as monthly measure of inflation and represents the change in the prices of a basket of goods and services.

4.3.3 Data filtering methodology

This section outlines the filtering methodology employed to derive sentiment scores from GDELT's Global Knowledge Graph (GKG) pertinent to economic growth and inflation, respectively. The filtering process involves three steps:

- Step 1: Keyword filter
- Step 2: Classification of observations with Bi-LSTM
- Step 3: Aggregation

First, a keyword-based thematic filter (economic growth, inflation) is applied to the themes field to identify news items relevant to the respective keywords. Second, the data from step one is further filtered on the themes field leveraging a neural network. Third, the filtered sentiment scores are aggregated to monthly frequency and country filters are applied.

An analysis of a random selection of 100 original news articles identified in step one reveals that the GDELT algorithm tends to recognise themes that do not actually exist. The algorithm's overzealousness warrants a more precise filter in order to reduce noise levels in the data set. In GDELT's GKG, every row represents one scanned news item. The themes field contains all themes the GDELT algorithm identifies as a string of labels, appearing in the same order they occur in the original text (Leetaru, 2015c). GKG themes include over 12,000 unique labels covering all aspects of human society.

A selection of 1,000 random articles is manually classified into relevant and non-relevant to a specific topic such as "economic growth" or "inflation" by referring to the original news article with the url contained in the DocumentIdentifier field. If the url is no longer available, the article is ignored.

Then, the GDELT themes extracted in step one are preprocessed. Each string of labels is divided into lower case tokens. The tokens for each news item are label-encoded – the themes are assigned numbers between zero and $N-1$. An "unknown" token is allocated to out-of-vocabulary words. I set a maximum length of 5,000 tokens and padding, standardising the length for each token sequence. The encoded themes serve as predictor and the binary classification into relevant and non-relevant observations serve as predicted information.

Model performance is assessed applying 10-fold cross-validation. In terms of metrics, performance is measured using precision, recall and the F1 score. Table 4.1 sets out the performance of a range of classifiers that were applied to the data set filtered for economic growth.

Table 4.1: Classifier performance

Classifier	Precision	Recall	F1
Gaussian Naïve Bayes	0.6747	0.5000	0.5744
Random Forest	0.8304	0.9118	0.8692
Support Vector Machine	0.8036	0.8645	0.8329
Unidirectional NN	0.8610	0.8649	0.8571
Bi-LSTM	0.8853	0.9375	0.9101

The Bi-LSTM classifier exhibits the strongest performance in terms of precision, recall and F1 metrics compared to the other algorithms that are explored.

I select a Bi-LSTM neural network due to its strong performance in terms of Precision and Recall among the range of models explored. The Bi-LSTM architecture created for the classification of themes includes two hidden LSTM layers, each incorporating 32 and 16 one cell memory blocks. Table 4.2 outlines the model architecture.

Table 4.2: Bi-LSTM model architecture

Layer type	Output shape
Masking	None, 5000
Embedding	None, 5000, 16
Bidirectional LSTM	None, 5000, 32
Bidirectional LSTM	None, 5000, 16
Dense	None, 5000, 8
Dropout	None, 5000, 8
Dense	None, 5000, 1

The "None" in the Output shape column stands for a number that is not predefined. By not assigning a specific value, the model has the flexibility to adapt this number as the batch size varies, deducing the shape from the context of the layers. The masking layer advises the classifier that some portion of the input is padding and should be disregarded. The output layer includes a single neuron to make predictions. It employs the sigmoid activation function to generate a probability output in the range of zero to one. A threshold of 0.5 is defined to transform the output to a discrete class (zero or one). Values below (above) this threshold are allocated to the first (second) class.

The sentiment scores obtained from the second step are filtered according to locations and aggregated to monthly frequency. The location field includes a list of all locations identified in an article (Leetaru, 2016).

4.3.4 The nature of sentiment scores

For each news item scanned, GDELT GKG's Tone field calculates a comma-delimited list of six emotional aspects, each represented as floating point number. From the Tone field, I use the average tone of a news article. This score takes values from -10 (very negative) to +10 (very positive), with zero being neutral (Leetaru, 2015a). The average tone score draws on the concept of sentiment mining, a method that counts words based on positive and negative pre-compiled dictionaries. The net sentiment stands for the overall tone (Hu and Liu, 2004).

Due to the duplication of GCAM emotion scores generated by different analysis methods, scores of four out of 24 analysis systems are selected to minimise overlap:

- WordNet-Affect was created by Strapparava and Valitutti (Strapparava and

Valitutti, 2004) extending the work on WordNet Domains (Magnini and Cavaglia, 2000). WordNet Domains link synsets, i.e. groupings of synonyms conveying the same idea, to domain labels such as Politics or Education. WordNet-Affect expands this concept in allocating emotional domain labels to the synsets. WordNet-Affect scores are word count-based and contain 280 emotional aspects such as “happiness”, “surprise” or “disgust”. For instance, occurrences of the word “happiness” in the original article will add to the “happiness” score.

- Loughran and McDonald explain that pre-compiled dictionaries for non-financial disciplines frequently misclassify words that have a financial context. The Financial Sentiment Dictionary employs negative word lists adapted to a financial setting to calculate word count-based scores (Loughran and McDonald, 2011). For instance, notions such as tax, cost, capital, and liability are usually not negative in financial documents but are classed negative by the Harvard dictionary. The method creates six scores.

- The Hedonometer scores generate a metric for societal happiness for English and several non-English languages (Dodds et al., 2011). In order to calculate the overall score, in excess of 10,000 unique words are scored by humans on a scale from one to nine, with five being neutral. Then, for each word, an average happiness rating is computed. For example, laughter, food and hate have been assigned scores of 8.5, 7.4 and 2.3. To calculate the happiness score of a news article, the scores for each word are averaged. The system generates 12 scores.

- ML-Senticon stands for a multi-layered synset-level lexicon and computes positivity and negativity scores for English and Spanish text (Cruz et al., 2014). As a first step, the synsets are given polarity scores. Subsequently, these scores are refined by building a graph of synsets, where the nodes are synsets. Edges between nodes are present if synset i is part of synset j. Then, a type of random-walk algorithm runs the positivity (negativity) scores over

> the edges of the graph to calculate the positivity (negativity) values for each synset. ML-SENTICON classifies synsets into eight successive levels, with each level containing more but lower-confidence synsets, permitting to optimise recall or precision when rating the synsets “positive” or “negative”. The system returns 32 scores.

See appendix A.1 for further information on the above sentiment scores.

The filtered observations are aggregated to monthly frequency. The mean and standard deviation of the average tone score is computed. Where the GCAM emotion scores are word count-based, the mean and standard deviation are computed and normalized to consider word count according to Baker *et al.* (Baker et al., 2016). For computed emotion scores, the monthly mean score and standard deviation are calculated. Further, the count of news articles and the total word count per month is returned.

4.3.5 The data sets

This section explains how the data sets used in this experiment are generated.

I apply the filtering methodology set out in section 4.3.3 to GDELT data to filter news articles for economic growth and inflation, respectively, and apply location filters for the 10 countries set out above.

I analyse the import and export volumes by trading partner to better understand the economic interconnectedness of each of the 10 countries (Datawheel et al., 2012). Six of the countries considered have trade connections to a variety of geographies. Poland, Norway and Turkey trade mainly with Western European countries. Over half of South Korea’s trade volume is linked to China. Given to the trade connections between countries, data relating to one economy may also affect another (Piccardi and Tajoli, 2018). Based on the idea of interconnection, a global data set containing data on all 10 countries is created for US, UK, Germany, Japan, Brazil and Mexico. A data set filtered for Western European locations is built for Poland, Norway and Turkey. For South Korea, a data set filtered for the location “China” is created.

The selection of countries considered in this experiment is evenly divided into developed and developing countries as per the MSCI Emerging Market Index (MSCI, 2021), diverse in size, determinants of economic growth and geography. For example, Germany is a developed eurozone country with an export-driven economy (cars, machinery) and diversified trade connections while Turkey is a developing country with strong trade links to Europe. Reliable macroeconomic data is available for all of the countries considered.

The filtered information is aggregated to monthly frequency, from March 2015 to June 2020.

Classifier predictions include true positives, false positives, true negatives and false negatives. The filtered data only retains true positive and false positive predictions, which equates to a noise level of c 5.4% and c 3.9% for the economic growth and the inflation filter, each.

For comparison, an unfiltered data sample of aggregated GCAM emotion scores is generated. For each year, around one million random observations are chosen and aggregated to monthly frequency. The unfiltered data set has a noise level of over 60%, meaning it contains news articles that are not relevant to the respective filters.

To control for macroeconomic effects, the baltic dry index and the crude oil price are included as explanatory variables when forecasting IP. The baltic dry index is indicative of global trade levels and is regarded a leading indicator for economic growth (Bildirici et al., 2015). A paper by Van Eyden *et al.* identifies a significant link between changes in oil price and economic activity in OECD countries (Van Eyden et al., 2019).

For models forecasting CPI, each countries' terms of trade indices and the crude oil price are added as explanatory variables. Mihailov *et al.* suggest that the expected relative change in the terms of trade is a key driver of inflation and more important than the domestic output gap (Mihailov et al., 2011). Salisu *et al.* describe a significant long-term positive connection between oil price and inflation (Salisu et al., 2017).

4.3.6 Data preprocessing

In this section, the data preprocessing steps are set out.

For each of the 10 aforementioned countries, the respective values for IP and CPI are used as predicted variables, describing the monthly percentage change.

For the above-mentioned variables, the augmented Dickey Fuller unit root test is applied to 20 years of monthly data and stationarity is not rejected at 5% significance.

For emotion scores only containing zeros, I assume that the GCAM system in question did not generate any scores and remove them. The scores impacted are predominantly based on the Hedonometer and ML Senticon methodologies. The GDELT data includes 664 raw scores and this preprocessing step reduces the amount of scores to 630, 628 and 632 for economic growth, inflation and unfiltered data sets, respectively. The monthly change in sentiment scores is computed. For the explanatory macroeconomic variables, the monthly change is calculated. The augmented Dickey Fuller unit root test is applied to the sentiment scores and explanatory macroeconomic variables and stationary is not rejected at 5% for any of them. Lastly, I standardize the sentiment scores and explanatory macroeconomic variables by removing the mean and scaling to unit variance.

4.4 Analysis approach

This section describes the analysis that is conducted to establish if the GDELT sentiment scores have predictive power for IP and CPI, respectively.

The Granger Causality between the GDELT sentiment scores and the predicted variables is examined to establish if the former variables are predictive of the latter ones (see section 3.2). I analyse the Granger Causality for lags up to three months and apply the Benjamini-Hochberg (BH) procedure to the generated p-values to adjust for multiple hypothesis testing (see section 3.3).

I look to incorporate a large number of narrative-based features derived from GDELT into an forecasting framework to predict IP and CPI, respectively. To achieve this, I employ a three-step process for creating a factor-augmented autoregressive forecasting framework as proposed by Girardi, Guardabascio and Ventura (Girardi et al., 2016).

First, for each of the economies considered, the predicted variable is forecast using an autoregressive model solely containing the predicted variable and explanatory macroeconomic variables.

It is not possible to incorporate the large number of variables extracted from GDELT in the autoregressive model described in Eq. (3.17). Hence, as a second step, I derive a small number of factors from the wide array of GDELT features for inclusion into the forecasting framework applying PLS as dimensionality reduction technique (see section 3.4).

PLS is applied to the residuals from step one. The residuals contain the portion of the predicted variable that is unexplained. Hence, implementing PLS on the GDELT features adds new information to the explanatory variables. The orthogonal relationship between the predicted variable and the residuals retains the orthogonality between the PLS components and the autoregressive components.

As a final step, the predicted variable, the explanatory variables and the PLS components are input into the autoregressive framework set out in Eq. (3.17) to create a factor augmented autoregressive framework (Colladon et al., 2019).

For performance comparison, three benchmarks are considered – first, an autoregressive model containing the predicted variable and explanatory macroeconomic variables, second, an autoregressive framework including the predicted variable, explanatory macroeconomic variables and factors derived from unfiltered GDELT sentiment and third, an autoregressive model containing the predicted variable, explanatory macroeconomic variables and standardized monthly change in the GDELT average tone score.

Performance is evaluated using walk-forward cross-validation (see section 3.6.3) and the root mean squared error (RMSE) metric (see section 3.6.2).

The modified Diebold Mariano test put forward by Harvey, Leybourne and Newbold is applied to establish whether model predictions are significantly different from benchmark predictions (see section 3.6.4).

4.5 Research findings

This section details the findings from the analysis that is conducted to establish if the GDELT sentiment scores have predictive power for IP and CPI, respectively.

4.5.1 Granger Causality test results

The Granger Causality test is applied to the GDELT sentiment scores and the macroeconomic indices for 10 countries with a maximum lag of three months. Tables 4.3 and 4.4 exhibit the number of BH-adjusted p-values that are significant at a level of 5% for each country's macroeconomic variable. The "Filtered" column stands for results from models containing GDELT scores filtered for economic growth and inflation respectively. The "Unfiltered" column displays the results for the models including the unfiltered GDELT scores.

Table 4.3: IP: Number of significant BH-adjusted p-values

Data set Country	Filtered	Unfiltered
US	5	0
UK	29	0
Germany	8	7
Norway	30	0
Poland	12	0
Turkey	8	1
Japan	6	0
South Korea	10	0
Brazil	35	16
Mexico	12	0

GCAM sentiment scores derived from filtered GDELT data consistently Granger-cause IP, while sentiment scores derived from unfiltered GDELT data show an inconsistent pattern, only Granger-causing IP for three out of 10 countries.

Table 4.4: CPI: Number of significant BH-adjusted p-values

Country \ Data set	Filtered	Unfiltered
US	14	0
UK	30	8
Germany	13	3
Norway	11	0
Poland	16	3
Turkey	57	1
Japan	19	0
South Korea	17	0
Brazil	39	0
Mexico	35	15

GCAM sentiment scores derived from filtered GDELT data consistently Granger-cause CPI, while sentiment scores derived from unfiltered GDELT data show an inconsistent pattern, only Granger-causing CPI for five out of 10 countries.

Results from the Granger Causality analysis suggest that a consistent number of filtered GDELT sentiment scores Granger-cause the respective macroeconomic indices. This in turn indicates that the filtering methodology proposed in section 4.3.3 has the capacity to generate sentiment scores that have a statistical relationship with IP and CPI.

Some reverse Granger Causality is present between macroeconomic variables and GDELT emotion scores albeit less consistent than set out in Tables 4.3 and 4.4. The findings support the notion that news sentiment Granger causes IP and CPI.

4.5.2 Forecast error analysis

From the respective filtered GDELT sentiment scores, three components are extracted applying PLS. These components are then combined with the explanatory macroeconomic variables to forecast IP and CPI, for 10 countries each with the model in Eq. (3.17).

The columns of Tables 4.5 and 4.6 set out the performance of the models incorporating filtered GDELT sentiment components compared to three benchmarks:

- an autoregressive model including only including predicted variable and explanatory macroeconomic variables (BM1);

- an autoregressive model containing predicted variable, explanatory macroeconomic variables and unfiltered GDELT sentiment factors (BM2);
- an autoregressive model including predicted variable, explanatory macroeconomic variables and the average tone score from GDELT (BM3).

The numbers in the cells stand for the RMSE in percentage terms for each model and its benchmarks. Blue (red) cells delineates outperformance (underperformance) of the models vs. the respective benchmarks. In the column “Sign.”, numbers in parentheses represent the number of significant coefficients related to GDELT components in the model in Eq. (3.17), with the asterisks alluding to the level of their statistical significance.

For instance, the first row in Table 4.5 illustrates that the model including the filtered US GDELT components beats all three benchmarks, with all three components being statistically significant (at 0.1, 0.05 and 0.01, respectively). In the case of Norway, instead, the model including filtered GDELT components only beats the second benchmark, with only one statistically significant component.

Table 4.5: Results of the model in Eq. (3.17) applied to IP.

Data set IP for	Model	BM1	BM2	BM3	Sign.
US	1.6348	1.6439	1.6527	1.6507	***(1),**(1), *(1)
UK	2.4663	2.4870	2.5065	2.4887	***(1)
Germany	2.7957	2.7978	2.7870	2.8032	**(2)
Norway	2.2907	2.2647	2.3410	2.2720	**(1)
Poland	7.6308	7.7321	7.8420	7.7351	***(1)
Turkey	4.5785	4.5836	4.6780	4.5843	**(2)
Japan	2.2138	2.2429	2.2571	2.2429	***(1), **(1)
South Korea	2.4776	2.5079	2.5579	2.5211	***(1)
Brazil	4.0484	4.0942	4.1341	4.0941	**(1)
Mexico	3.2630	3.2494	3.2532	3.2471	***(1)

Numbers represent the RMSE (%). Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (3.17) ($***$ denotes at least one GDELT sentiment factor with p-value < 0.01, $** < 0.05$, $* < 0.1$).
Eight out of ten models containing filtered GDELT sentiment factors outperform their benchmarks. All models include at least one statistically significant GDELT sentiment factor.

The modified Diebold Mariano test demonstrates that model predictions for IP are statistically different from BM1 and BM2 in nine out of ten, and for BM3 in 10 cases, respectively at 1, 5 or 10% significance (see Table A.1).

Due to country-wide lockdowns in many parts of the world that severely affected economic activity, IP metrics for all ten economies considered display elevated levels of volatility during the first half of 2020. Considering the cross-validation data sets, the last validation set covers the COVID-19 pandemic in 2020. Across countries, forecasts on this validation data exhibit a considerably larger RMSE than forecasts on validation sets that are unaffected by the pandemic. Nonetheless, performance dynamics on the last validation set remain unchanged in that most models including the GDELT factors beat their benchmarks.

Table 4.6: Results of the model in Eq. (3.17) applied to CPI.

Data set CPI for	Model	BM1	BM2	BM3	Sign.
US	0.2051	0.2031	0.2165	0.2038	***(1)
UK	0.3212	0.3118	0.3258	0.3193	***(2)
Germany	0.4117	0.4316	0.4288	0.4279	***(1)
Norway	0.4715	0.4727	0.4706	0.4640	
Poland	0.3321	0.3383	0.3649	0.3512	***(1)
Turkey	1.0194	1.0263	1.0513	1.0685	***(2), **(1)
Japan	0.2464	0.2485	0.2538	0.2524	***(2)
South Korea	0.3941	0.4055	0.4057	0.4070	**(1)
Brazil	0.3876	0.3993	0.4165	0.3946	***(1)
Mexico	0.4226	0.4349	0.4340	0.4581	***(1)

Numbers represent the RMSE (%). Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (3.17) ($***$ denotes at least one GDELT sentiment factor with p-value < 0.01, $** < 0.05$, $* < 0.1$).

The models containing filtered GDELT sentiment factors beat BM1 for eight, BM2 for nine and BM3 for seven out of ten economies, respectively. Nine out of ten models include at least one statistically significant GDELT sentiment component.

The adjusted Diebold Mariano test statistics indicate that model predictions for CPI are different from BM1 for all and for BM2 and BM3 for seven out of ten countries, respectively at either 1, 5 or 10% significance (see Table A.2).

According to the findings from the forecast error analysis, those models using filtered sentiment perform consistently better that those using unfiltered sentiment, suggesting that the filtering methodology isolates relevant signals. The results demonstrate that GDELT sentiment scores improve forecasts for IP and CPI for most economies considered. They indicate that adding a wide spectrum of emotions to model inputs generates better predictions compared to only incorporating one single sentiment score such as the average tone.

4.5.3 Drivers of GDELT factors

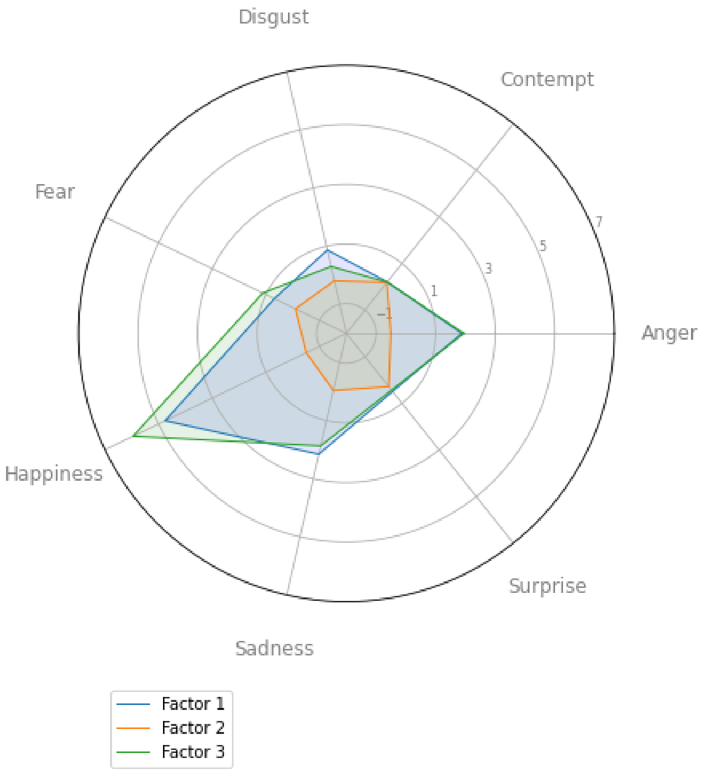
In order to gain an understanding of the relationship between raw GDELT sentiment scores and the PLS components extracted from the filtered GDELT data, each component's loadings are analysed.

Loadings represent the strength of relationship between the raw sentiment scores and the PLS components, quantifying the pertinence of the original sentiment scores in each of the components.

All sentiment scores from GDELT correspond to a specific emotion such as "euphoria", "love" or "joy" and are manually associated with seven universal emotion groups as proposed by Ekman and Corduro (Ekman and Corduro, 2011). These seven emotion groups describe emotions as discrete, automatic reactions to events and state that emotions such as happiness or anger describe categories of associated states with specific shared characteristics. According to these seven categories, the aforementioned emotions are mapped to "happiness".

For each PLS component, the loadings are summed up according to these seven emotion groups. Associating raw sentiment scores with emotion groups lends some interpretability to my analysis as this provides insights into the strength of relation between distinct emotion groups and each PLS component.

Figure 4.3: IP: Significant PLS components explained by emotions (US)

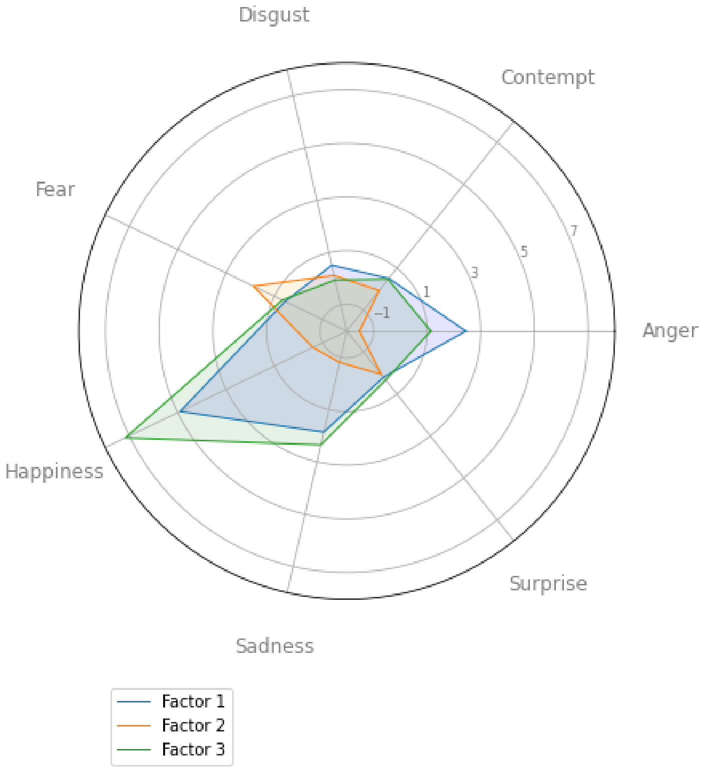


Factor 1 and Factor 3 are have the strongest relationship with sentiment scores belonging to "happiness", while Factor 2 has the strongest link to emotion categories "surprise" and "contempt".

Figs. 4.3 and 4.4 show radar charts of the emotions related to the loadings of the statistically significant PLS components used to predict IP and CPI in the US and Turkey, respectively. As can be seen from Tables 4.5 and 4.6, the corresponding models beat all three respective benchmarks and are associated with substantial statistical significance. Further charts can be provided upon request.

This example illustrates that the components incorporated into forecasts of IP and CPI can be linked to well defined emotions. For this reason, changes in such emotions – as expressed in news articles published by global newspapers – help to explain changes in major macroeconomic indicators. Of the seven distinct emotion groups, "happiness" and "anger" exhibit the strongest predictive power, both displaying a positive relationship with the respective component. This is the case for all PLS factors.

Figure 4.4: CPI: Significant PLS components explained by emotions (Turkey)



Factor 1 and Factor 3 are have the strongest relationship with sentiment scores belonging to “happiness”, while Factor 2 has the strongest link to emotion categories “fear”, “contempt” and “surprise”.

4.6 Discussion

This experiment sets out a new method of including emotions from global newspapers into forecasts of macroeconomic variables. It proposes a filtering methodology to derive and aggregate large amounts of information, which is implemented to create data sets filtered for economic growth and inflation. For 10 diverse economies, IP and CPI are forecast, considering each country’s trade connections when applying location filters. The filtered data reveals consistent Granger Causality for IP and CPI. Autoregressive models containing the filtered GDELT data beat their benchmarks for most predicted variables. In particular, models including a wide array of emotion scores generally outperform those models solely incorporating one aggregate sentiment score, average tone. Based on these results, $H1$ cannot be rejected. Associating raw GDELT sentiment scores with specific emotion groups provides insights into the relationship of these groups with each PLS component, indicating that “happiness” and “anger” are their main drivers.

Chapter 5

Macroeconomic forecasting with statistically validated knowledge graphs

5.1 Introduction, motivation and background

Stories are an integral part of society. A rapidly evolving body of literature explores narrative as driver of human actions and examines how these actions are ultimately reflected in economic outcomes (see chapter 2 for details).

So-called *knowledge graphs* are one way to capture features from narrative (e.g. topics) as well as the interaction of such features. A broad range of knowledge graph applications has shown that graph modeling and analytics capture complex relationships well (Emmert-Streib et al., 2018).

This experiment leverages research on economic knowledge graphs to create a graph-based methodology for macroeconomic forecasts. My approach uses machine learning, graph analytics and natural language processing to extract themes from global newspaper text and build theme-based graphs that reflect the state of the economy. I demonstrate how features derived from such graphs can be incorporated into forecasting frameworks to improve predictions of industrial production (IP) in three large countries. Results are complemented by an interpretability study illustrating that disease and economy-related themes have the strongest predictive capacity.

5.2 Literature review

This section discusses a choice of existing research on knowledge graphs and their approach to the study of economic systems. Further, this section covers literature on the extraction and analysis techniques most relevant to weighted, undirected graphs as they are the object of this experiment. Lastly, I formulate the research questions this project attempts to answer and set out the experiment's contribution to literature.

The interpretation and analysis of economic systems in terms of knowledge graphs has enjoyed increasing attention in recent years across disciplines such as economics, the social sciences, computer science, statistics, business and management (Emmert-Streib et al., 2018).

5.2.1 Economic and financial graphs

Existing works focus on a variety of economic knowledge graphs. For example, an analysis of a graph of interbank spreads establishes that they are significantly impacted by graph centrality metrics, in particular during the global financial crisis in 2008 (Temizsoy et al., 2017). Using a knowledge graph of European banks, Constantin *et al.* examine the effect of negative shocks on bank returns, revealing an interconnectedness of bank returns that could be used to create an early warning framework for bank distress (Constantin et al., 2018). An analysis of supply chain disruptions after the Japanese tsunami in 2001 suggests that the interruption of input-output connections among firms resulted in a 1.2% decrease in the country's gross output in 2012 (Carvalho et al., 2021). Adamic *et al.* leverages graph analytics to delineate the time-series aspects of data and liquidity flows in the E-mini S&P stock index futures market, demonstrating strong simultaneous relationships between graph features and financial variables, with the former Granger-causing liquidity metrics such as intertrade duration, and trading volume (Adamic et al., 2017). A paper by Piccardi and Tajoli suggests that complex products have a very central position in the global trade graph, resulting in an unbalanced distribution of trade connections between geographies and increased sensitivity to specific shocks that can be quantified using graph analytics (Piccardi and Tajoli, 2018). Guo and Vargo outline that population, trade, cultural proximity, and geographic closeness affect international news attention using a graph of themes and location data from GDELT (Guo and Vargo, 2020). Campi *et al.* examine the drivers of specialisation in agricultural production, connecting countries to their agricultural products via temporal bipartite graphs (Campi et al., 2020). The authors find that agricultural production can be represented as a dense graph of well-defined communities of economies and products that are determined by environmental, economic, socio-political and technological factors. Bellomarini *et al.* use a network of business relationships between Italian companies to spot transactions that may result in strategic company takeovers (Bellomarini et al., 2020). Gomez *et al.* leverage graph theory to interpret business cycle synchronization in the EU over the last 20 years, illustrating that co-movements and interactions between economies have been relatively stable before the Eurozone crisis in 2011/12 and have become more frequent thereafter (Matesanz Gomez et al., 2017).

5.2.2 Macroeconomic forecasting with knowledge graphs

Only a few studies have applied knowledge graphs to macroeconomic analysis and forecasting. For example, a paper by Bonaccorsi *et al.* assesses country centrality in multilayer graphs, showing that such metrics reflect the global North-South divide and have a positive relationship with economic indicators such as GDP (Bonaccorsi et al., 2019). Yang *et al.* derive entities from news text and connect them to macroeconomic indicators in knowledge graphs. The authors find that the inclusion of graph metrics in a macroeconomic forecasting framework significantly enhances forecasts of inflation, net export growth, and housing prices (Yang et al., 2020).

5.2.3 Graph filtering

The analysis of knowledge graphs is an effective tool to interpret and examine complex systems. Nonetheless, real-world graph data sets can be very large, making it hard to recognise those components that are most relevant to such systems (a graph's "backbone"). Research sets out several methods for extracting graph backbones, which are discussed in section 3.7.3. By design, graph modelling is parsimonious and hence, condensing news text into theme-based graphs can potentially enhance traditional macroeconomic forecasting models. Combining macroeconomic forecasting and knowledge graph theory could improve the monitoring of global economic developments. This could be pertinent to organisations that base far-ranging decisions on their macroeconomic forecasts, such as central banks or supranational institutions such as the UN.

5.2.4 Research questions and hypotheses

In this experiment, I attempt to predict industrial production (IP) with data from GDELT. I look to answer two research questions, i.e., (1) whether socio-economic interactions can be reflected in knowledge graphs based on GDELT themes, and (2) whether features derived from such graphs are predictive of changes in economic activity. Accordingly, I formulate the two following hypotheses:

- H_1: Changes in the structure of graphs based on GDELT themes are predictive of socio-economic changes.
- H_2: Features extracted from such graphs enhance forecasts of economic activity.

By addressing these questions, I contribute to research on economic graphs in at least three aspects. First, I show how themes from news text can be included into macroeconomic forecasting frameworks to enhance predictions of economic activity. Second, I advance the literature on economic graphs by introducing a data-driven approach to operationalise tasks such as information extraction, feature generation and macroeconomic forecasting. I supplement my findings with an interpretability analysis illustrating that disease and economy-related themes have the strongest predictive capacity. Third, I add to literature on monitoring socio-economic changes through news narrative by analysing the development of graph metrics over time. Narrative-based knowledge graphs can be adapted to track a variety of phenomena such as "climate change", "migration" or "human rights violations" in real-time and therefore have many use cases beyond macroeconomic forecasting.

5.3 Data and methods

This section uses GDELT as data source (see section 4.3.1 for details) and describes the filtering methodology applied to obtain relevant signals.

The GDELT algorithm returns the identified themes for each news item as a string of labels. These themes are extremely detailed and can be classed into distinct groups such as "economic", "government" or "health" (see appendix C.1 for a list of all theme categories).

Theme groups represent events or conditions, except for four solely descriptive categories ("actor", "ethnicity", "language" and "animal"), which are therefore excluded. Table 5.1 sets out the number of themes for each of the three countries considered in this experiment, both in the raw data (no. of themes) and the number of themes after excluding the purely descriptive ones (reduced no. of themes).

Table 5.1: Number of of GDELT themes

Country	no. of themes	reduced no. of themes
US	6880	3501
Germany	5313	2925
Japan	3330	1994

Removing purely descriptive themes belonging to categories "actor", "ethnicity", "language" and "animal" significantly reduces the overall count of distinct themes as these groups are very nuanced.

5.3.1 Predicted variables

This experiment attempts to forecast industrial production (IP) for the US, Germany and Japan. IP is a monthly indicator of economic activity, monitoring the monthly change in the volume of production output across a variety of organisations. Germany, the US and Japan are chosen for representing large, industrialised economies with diverse trade links in three regions of the globe – America, Europe and Asia.

5.3.2 Filtering methodology

I leverage the filtering methodology outlined in section 4.3.3 to extract news articles relevant to economic growth from GDELT data (see 4.3.1 for details). Then, I apply country filters for the US, Germany and Japan, respectively to the locations field in the filtered data set. For each country and each calendar month, I extract co-occuring theme pairs and the count of their co-occurrences from GDELT GKG. This is achieved by a query applied via BigQuery to the filtered data set that is stored on the Google Cloud Platform. The query is applied to each calendar month and combines several steps:

- Split string of themes into unique tokens;
- Extract all combinations of theme tokens that occur in the same article;
- Count co-occurrences of theme pairs. Pairs with identical themes are disregarded;
- Apply location filter for the US, Germany and Japan, respectively.

I then construct theme-based undirected weighted knowledge graphs for each calendar month from March 2015 to December 2020. In these graphs, the themes are nodes, the co-occurrences of themes are edges and the count of the co-occurrences are edge weights. I create the same knowledge graphs using unfiltered GDELT themes for benchmarking purposes.

5.3.3 Statistical validation of graph elements

In order to obtain the backbone of each monthly theme-based graph, I apply the disparity filter to identify statistically relevant pairs of edges, referred to as the "backbone".

The monthly graphs are connected and very large with c 4,000 nodes and c 1 million edges. To control for multiple hypothesis testing, the p-values are adjusted applying the Benjamini-Hochberg (BH) procedure (Benjamini and Yekutieli, 2005). Each edge has to be tested twice since the normalisation of edge weights is carried out with regard to one of the two nodes they are associated with, which means that the null hypothesis is at least rejected once for those edges preserved in the backbone.

Implementing the disparity filter on the monthly theme-based graphs returns connected graphs with c 50% and c 90% of the original number of nodes and edges, respectively.

5.3.4 Graph features

This section describes the computation of features from monthly theme-based knowledge graphs. Details on the derivation of these metrics can be found in section 3.7.

First, the eigenvector centrality is calculated for all nodes, for each monthly graph, which measures a node's importance in a graph structure.

A $T \times K$ matrix is built, where T describes the number of calendar months (i.e. observations) and K represents the number of nodes (i.e. themes). Specifically, the matrix incorporates an eigenvector centrality score for each node and each calendar month.

As a second metric, the portrait divergence for each graph compared to the previous month's graph is computed. This time series includes one portrait divergence score for each calendar month.

5.3.5 Explanatory variables

When modeling IP, the monthly changes in the baltic dry index and the crude oil price are included to account for macroeconomic effects.

The baltic dry index delineates global trade volume and is regarded a leading indicator for economic activity (Bildirici et al., 2015). Van Eyden *et al.* establish a strong connection between oil price fluctuations and economic growth in OECD countries (Van Eyden et al., 2019).

5.3.6 Data preprocessing

This section describes the data preparation methods.

In this experiment, I am looking to predict the monthly changes in IP indicators for the US, Germany and Japan. To check the data for stationarity, the augmented Dickey Fuller unit root test is implemented on 20 years of monthly data for the IP indicators and the explanatory variables; and stationarity is not rejected for any of them at 5% significance.

Any missing entries in the $T \times K$ eigenvector centralities matrix are replaced by zero, the lowest possible centrality score. The augmented Dickey Fuller unit root test is carried out for the eigenvector centralities and portrait divergences scores and stationary is not rejected at 5% for any of these metrics. The predictive variables are normalised by removing the mean and scaling to unit variance.

5.4 Analysis approach

This section outlines the analysis carried out to gauge if the narrative-based graph features have predictive capacity.

I analyse the Granger Causality for lags up to three months (see section 3.2). The Benjamini-Hochberg (BH) procedure is applied to the resulting p-values to adjust for multiple hypothesis testing (see section 3.3).

I forecast IP for the three aforementioned economies using the factor-enhanced autoregressive framework outlined in section 4.4.

I compare model performance to three benchmarks – first, an autoregressive model only incorporating the predicted variable and explanatory macroeconomic variables, second, an autoregressive model including the predicted variable, explanatory macroeconomic variables and each country's portrait divergence scores, and third, an autoregressive model inputting the predicted variable, explanatory macroeconomic variables and five PLS factors derived from eigenvector centralities derived from unfiltered GDELT data.

Model performance is evaluated applying walk-forward cross-validation (see section 3.6.3) and the root mean squared error (RMSE) (see section 3.6.2). I implement the modified Diebold Mariano test to check whether model predictions are significantly different from benchmark predictions (see section 3.6.4).

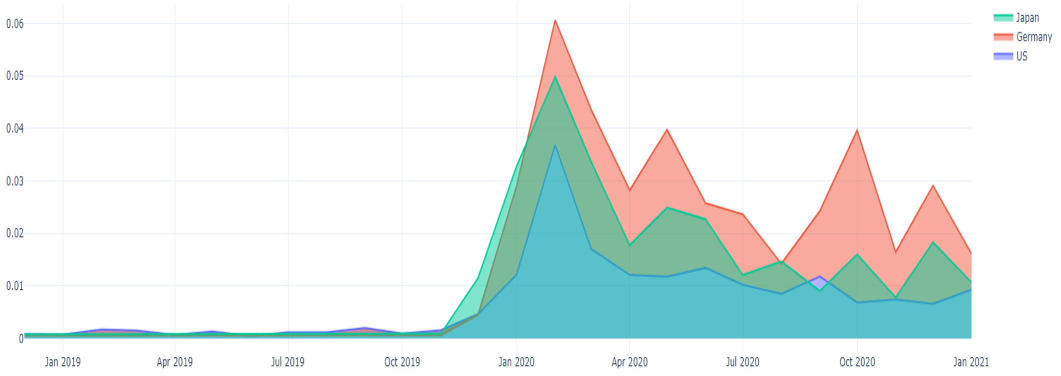
5.5 Research findings

This section discusses the findings from the analysis outlined in the previous section.

5.5.1 Evolution of graph features

Theme-based knowledge graphs can be used to spot changes in specific aspects of social systems, exemplified here by the evolution of the COVID-19 outbreak. Fig 5.1 illustrates the development of the median eigenvector centrality score of COVID-19 symptom related themes. These themes were chosen manually for their pertinence, such as “pneumonia”, “fever” or “cough”, within monthly graphs for the each of the three economies considered (see section B.1 for a complete list of themes applied). The example demonstrates that the evolution of the monthly median eigenvector centralities is consistent with the timing and impact of events in specific geographies.

Figure 5.1: Influence of COVID-19 related themes within monthly graphs.



Before the end of 2019, the median eigenvector centralities related to COVID-19 symptoms were very low for all three economies, then increasing considerably as the pandemic spread around the world, first reaching Japan, then Germany and finally the US.

5.5.2 Granger Causality test results

I implement the Granger Causality test on the eigenvector centralities from theme-based graphs and the IP index values with lag of up to three months. Table 5.2 displays the amount of BH-adjusted p-values that are statistically significance at 5% for each economy's IP index.

Table 5.2: IP: Number of significant BH-adjusted p-values

Data set Country	GC (filt)	Reverse GC	GC (unfilt)
US	673	583	94
Germany	783	742	67
Japan	552	556	99

GC refers to Granger Causality; filt (unfilt) refers to filtered (unfiltered) GDELT data.
The Granger Causality test applied to features created from filtered GDELT data results in a considerably higher number of statistically significant BH-adjusted p-values, compared to the same test applied to features generated from unfiltered GDELT data.

The results from Table 5.2 suggest that the filtering methodology set out in in section 5.3.2 effectively identifies relevant signals. They also indicate the existence of strong reverse Granger Causality between IP and eigenvector centralities. Due to the two-way nature of news reporting, this is unsurprising; both events (identified by GDELT as themes) and IP are covered and commented on in the press.

5.5.3 Forecast error analysis

For the US, Germany and Japan, the eigenvector centrality matrices are distilled into five components by implementing PLS. They are then, together with the explanatory variables, input into the forecasting framework outlined in section 4.4 to predict IP. According to BIC and AIC, all models have a one month lag.

The columns of Table 5.3 display the performance metric (RMSE) from the factor enhanced models measured against three benchmarks. These benchmarks are autoregressive models containing different sets of predictors as described in section 5.4. The numbers in the cells stands for the RMSE (in %) for each model and its benchmarks. Blue (red) cells conveys model outperformance (underperformance) against the benchmarks. The numbers in the column “Sign.”, represent the number of significant coefficients related to GDELT components in the forecasting framework in Eq. (3.17), with the asterisks describing the level of their statistical significance.

Table 5.3: Results of the model in Eq. (3.17) applied to IP.

IP for	Model	BM1	BM2	BM3	Sign.
US	1.6139	1.6341	1.6376	1.6144	***(1), **(1)
Germany	2.6942	2.7079	2.7082	2.7159	*(1)
Japan	2.4978	2.5300	2.5321	2.5049	*(1)

Numbers represent the RMSE (%). Blue (red) cells denote cases in which the model outperforms (underperforms) the benchmark. Numbers in parentheses correspond to the number of significant coefficients associated with GDELT factors in the model in Eq. (3.17) ($***$ denotes at least one GDELT sentiment factor with p-value < 0.01, $** < 0.05$, $* < 0.1$).
Rows one to three in Table 5.3 illustrate that the models including the filtered GDELT components beat all three benchmarks for the US, Germany and Japan, respectively. All factor augmented models have at least one statistically significant GDELT component at 10%, 5% or 1%.

The last validation set in cross-validation covers the COVID-19 outbreak in 2020 and forecasts on this last validation set show a considerably higher RMSE across countries compared to forecasts on the other validation sets.

Table B.1 shows the p-values from the modified Diebold Mariano test, which is applied to establish if factor enhanced model predictions are significantly different from the benchmark forecasts as outlined in section 5.5.3.

The statistics in Table B.1 show that all model forecasts for IP are statistically different to BM1, BM2 and BM3 forecasts at significance levels of 1 % or 10 %.

Findings indicate that features from theme-based graphs enhance predictions of IP for three large economies. The factor augmented autoregressive models consistently outperform those frameworks including a single graph similarity measure (BM2) or components derived from unfiltered GDELT data (BM3). This suggests that eigenvector centrality scores contain valuable information and are therefore appropriate for capturing the changes in the structure of theme-based knowledge graphs.

5.5.4 Drivers of GDELT factors

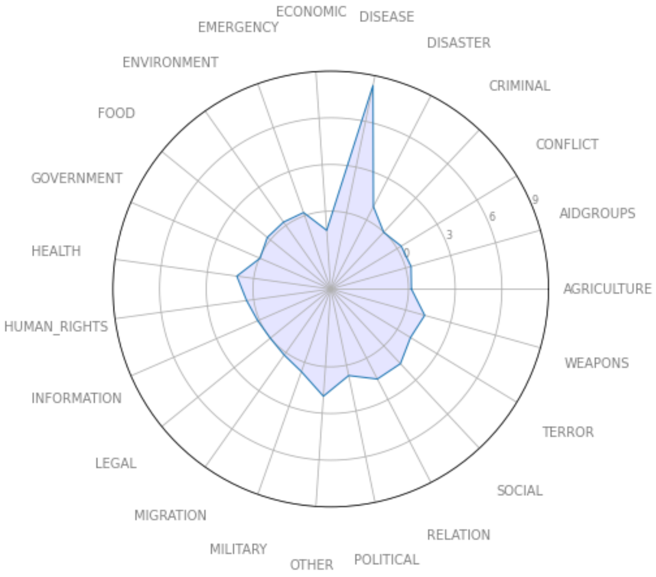
In this section, I evaluate the loadings associated with each PLS component to get insights into the link between themes and those components. Loadings describe the strength of connection between the eigenvector centrality scores for each graph node (i.e. GDELT theme) and the PLS components, measuring the influence of the GDELT themes in each of them.

GDELT themes describe events or conditions such as "civil unrest", "food shortage" or "terrorist attack" and are assigned to 23 theme groups using the GDELT naming convention. For instance, themes including the word "environment" are mapped to the environment category; themes including the word "food" are mapped to the food category, etc (see section C.1 for detailed list).

For each component, the loadings are added up according to the 23 theme groups outlined above. Assigning GDELT themes to theme groups provides an understanding into the relationship of these groups with each PLS component.

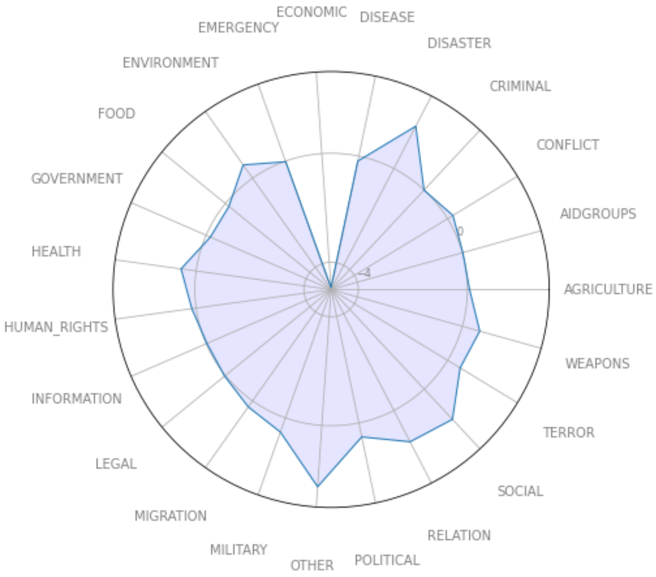
The radar charts in Figs. 5.2 and 5.3 depict the theme groups related to the loadings of the statistically significant PLS components used to predict IP in the US. Table 5.3 shows that the model in question beats its three benchmarks and displays robust statistical significance.

Figure 5.2: Significant PLS components explained by theme categories (Factor 1, US)



Factor 1 relating to the US model has a strong positive relationship with theme category "disease".

Figure 5.3: Significant PLS components explained by theme categories (Factor 4, US)



Factor 4 relating to the US model has a strong negative relationship with theme category "economic".

This example suggests that the components we use to model IP can be associated to specific theme groups. Over time, fluctuations in such groups help explain changes in IP, which is regarded as a monthly proxy of economic activity.

“Disease” and “economic” theme groups have the strongest relationships with component 1 and component 4, respectively and hence can be regarded as key drivers of these components with the strongest predictive capacity. This holds true for the statistically significant PLS components for the models forecasting German and Japanese IP.

5.6 Discussion

This experiment proposes a novel approach of including themes from global newspaper text into macroeconomic forecasts. I implement an effective filtering methodology for isolating pertinent signals from a large amounts of data. The filtered data is employed to create monthly theme-based weighted undirected knowledge graphs for for the US, Germany and Japan, respectively. For benchmarking, monthly knowledge graphs are built using unfiltered GDELT data. For each graph, I compute the eigenvector centrality and portrait divergence scores, effectively creating a monthly frequency time series of variables. I demonstrate that the change in graph centralities is associated with genuine real-world change in socio-economic systems. Based on this finding, I cannot reject hypothesis H_1.

Results from the Granger Causality test suggest that a strong statistical relationship exists between graph features and each country’s IP values. For each economy considered, I apply PLS to compress the eigenvector centralities into five components and incorporate them, together with the explanatory variables, into a factor augmented autoregressive model. Findings reveal that variables from theme-based graphs significantly improve predictions of IP for each of the three countries considered. Notably, the factor augmented models consistently beat those benchmarks that only include a single graph similarity metric (portrait divergence) and those that incorporate five PLS components based on unfiltered GDELT data. Based on these results, I cannot reject hypothesis H_2. Assigning over one thousand GDELT themes to 23 specific theme groups lends some interpretability to the relationship between those groups and each PLS component, with themes related to “disease” and “economic” being the components’ key drivers.

Chapter 6

Predicting market inflation expectations with news topics and sentiment

6.1 Introduction

In this chapter, I apply the methodologies and data developed in previous chapters to a new domain, namely the prediction of market-based inflation expectations.
My experiment advances existing literature by including themes and sentiment from global news narrative into predictions of breakeven inflation rates (BEIR) movements leveraging the Global Database of Events, Language and Tone (GDELT) as data source (Leetaru, 2015b). This database has – to the best of my knowledge – not yet been applied to this area of research. Most publications in this field focus on the US. My experiment takes a much broader, global approach by covering eight economies with mature fixed income markets around the globe. I supplement my findings with a feature importance analysis, indicating that economic and financial themes have the strongest predictive capacity in the predictions I make. I analyse the impact of cross-country spillovers of news narrative on BEIR movements using Graphical Granger Causality modelling and graph analytics, which can be found in the US and Germany, while local narrative affects BEIR movements in five other

economies considered.

6.2 Literature review

This section covers a choice of existing publications on the effect of news reporting on inflation expectations.

Most of the works in this area quantify consumer inflation expectations through surveys. These papers rely primarily on regression models to forecast monthly data. A lot of studies focusing on the US quantify inflation expectations using the University of Michigan Survey of Consumers and overall concur that newspaper reporting influences a change in inflation expectations.

Doms and Morin (Doms and Morin, 2004) measure the sentiment and volume of economic news using The Economist's recession indicator. The authors find that US consumers change their expectations more often during times of increased news coverage. This finding supports the theory of consumer expectations stickiness proposed by Sims (Sims, 2003), which explains why the tone and volume of economic news coverage affects consumers beyond the economic information contained in news articles. Pfajfar and Santoro (Pfajfar and Santoro, 2013) analyse the link between press coverage on inflation and US consumers' inflation expectations, applying an expectation formation theory based on an epidemiological concept (Mankiw and Reis, 2003; Carroll, 2003), according to which consumer expectations are affected by news narrative (presumed to contain professionals' projections). According to this study, consumers are more open to negative news, with a general disconnect between the accuracy of consumers inflation expectations, news flow on inflation and regularity of expectation updating. Similarly, Dräger and Lamla (Dräger and Lamla, 2017) examine the drivers of changes in consumers' inflation expectations. They conclude that news stories lead to changes in expectations that improve in accuracy. This finding supports the concept of imperfect information (Woodford, 2001; Sims, 2003; Mankiw and Reis, 2003). A paper by Larsen *et al.* (Larsen et al., 2021) evaluates the impact of the media reporting on consumers' inflation expectations formation. Results reveal that the topics covered by media outlets have indeed predictive capacity for inflation expectations. The degree of consumer information rigidity fluctuates due to changing volumes of pertinent press coverage, triggering an adjustment in consumer inflation expectations. D'Acunto *et al.* (D'Acunto et al., 2019) investigate individual consumers inflation expectations based on grocery shopping, gathering information through surveys. According to this study, price changes in groceries are interpreted as "news", with consumers' expectations being highly sensitive to price changes in products that they purchase often. Taking a different approach, Mazumder (Mazumder, 2021) analyses on the number of "Fed" mentions in US press reporting and concludes that this information improves the modelling of US inflation expectations.

Similarly to works focusing on the US, research on the effect of news reporting on European inflation expectations come to comparable conclusions. For example, a study by Jansen and Neuenkirch (Jansen and Neuenkirch, 2017) gathers information from Dutch households via surveys to investigate the connection of inflation perceptions and newspaper readership. The authors reveal that the recognition of price fluctuations is a key determinant of the precision of one-year-ahead inflation expectations, with increased newspaper consumption resulting in more precise expectations. Lamla and Lein (Lamla and Lein, 2014) examine the effect of newspaper stories covering inflation-related themes on German consumers' inflation expectation. Findings indicate that the amount and tone related to inflation-related news have a causal relationship with consumer inflation expectation formation.

While many of the papers in this field employ monthly frequency consumer survey data as proxy for inflation expectations, a few works use market-based inflation expectation measures such as inflation breakeven rates or interest rate swaps, proposing that news topics and sentiment have a causal connection with financial markets and economic indicators. For example, Bauer (Bauer, 2015) uses market-based measures of inflation expectations such as US inflation breakeven rates and US interest rate swaps and finds that they react to macroeconomic news. Bybee *et al.* (Bybee et al., 2020) extract topics from US business news text via natural language processing techniques and use them to build attention indices. These indices are then incorporated in 1,000 rolling day lasso regressions to model interest rate swaps. A coefficient analysis shows that over time, model performance is driven by a range of different topics. A paper by Kabiri *et al.* (Kabiri et al., 2020) examines the effect of news sentiment on the US economy and US capital markets. Results show that abrupt changes in sentiment – known as "shocks" – have a clear impact on economic activity and market-based indices such as the S&P 500 index, credit spreads and interest rates.

In this experiment, I look to predict short-term movements in 10 year BEIR for eight diverse economies, using data from the Global Database of Events, Language and Tone (GDELT) and tackle two research questions, i.e., (1) whether themes from GDELT and their associated average tone have predictive capacity in regard to short-term fluctuations in BEIR, and (2) whether news reporting has cross-country spillover effects on BEIR. Hence, I state the two hypotheses:

- H_1: GDELT themes and their associated average tone enhance predictions of short-term changes in BEIR.
- H_2: News reporting has cross-country effects on BEIR.

By addressing the above questions, I contribute to research on the forecasting of market-based inflation expectations in at least three aspects. First, I show how topics and their associated sentiment from newspaper stories can be included into machine learning frameworks to enhance predictions of BEIR for eight diverse economies. Second, I advance literature on predicting inflation expectations by putting forward a data-driven approach to operationalise concepts such as information filtering, feature creation and predictive modelling. I supplement my findings with a feature importance analysis illustrating that economy-related themes have the strongest predictive capacity. Third, I use Graphical Granger Causality modelling and knowledge graph analytics to advance the work on the cross-country impact of news reporting on BEIR .

6.3 Data and methods

This section sets out the data and variables used in this experiment.
As data source, I use the theme and average tone fields from the GDELT GKG (see 4.3.1 for details).

6.3.1 Predicted variables

I look to to forecast short-term changes in 10 year BEIR for eight diverse economies with mature inflation-linked bond markets – the US, the UK, Germany, Japan, South Africa, Australia, Brazil and Mexico. The BEIR is the difference between nominal and inflation linked bond yields and is used as an effective indicator of market-based inflation expectations that are reflected in the market (Pimco, 2021).

6.3.2 Explanatory variables

This experiment models daily frequency data, unlike most of the works in this area. Market-based measures of inflation expectations are very receptive to data surprises, both in terms of stock market prices and news reporting (Bauer, 2015). I employ market-based variables to account for the effect of stock market conditions such as investors' risk perceptions and economic outlook on BEIR (Ciccarelli and Garcia, 2009). Further, I include commodities, as existing works find a strong connection between fluctuations in commodity prices and movements in BEIR (BIS, 2011; da Cunha Cabral et al., 2021). Therefore, I include the following variables to account for the described relationships:

- Each country's stock market index;
- Each country's FX rate;
- Each country's yield curve steepener;
- Gold price;
- Crude oil price;

- Bloomberg Commodity Total Return index.

In reference to hypothesis H_1, I state that market-based variables include information that is already factored into BEIR, while variables extracted from news text hold information not yet priced into the markets and therefore enhance predictions of BEIR fluctuations. In order to account for such data surprises, I incorporate features derived from GDELT themes and their associated average tone score into predictions of BEIR movements.

6.3.3 GDELT data sets

This section sets out the filtering method applied to GDELT data for building the data sets used in this experiment.

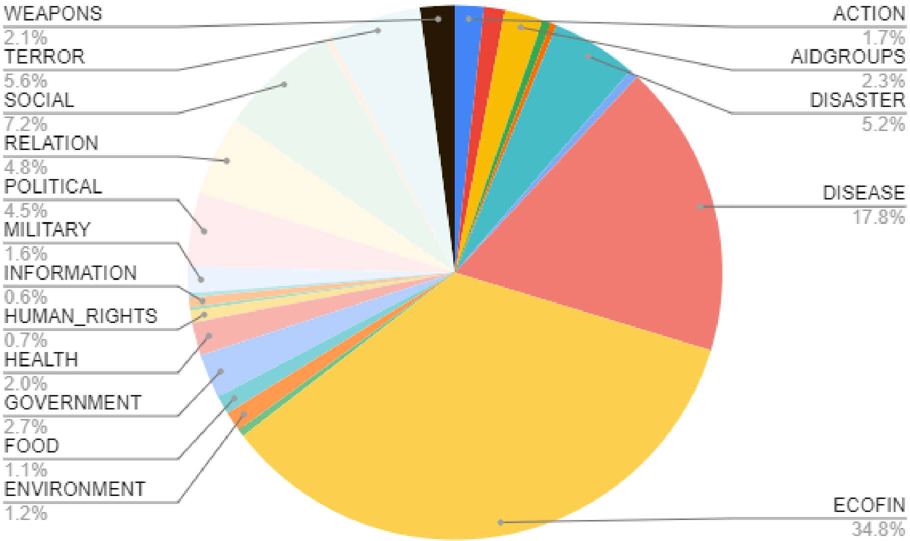
I extract GDELT themes that represent the economic and financial environment for each of the eight economies considered in this experiment. To accomplish this, I leverage the GDELT theme taxonomy, filtering for themes with the prefix "ECON". Therefore, in order to identify pertinent surprises as per H_1, I only retain observations that contain at least three themes related to finance or the economy. As news are always reported in context, the themes are expected to also include topics related to areas other than finance and the economy. This data set is stored on the Google Cloud Platform, accessible via Google BigQuery. Then, the themes and their associated average tone score are extracted for each of the eight economies considered. This is done with a query that combines several steps:

- Split string of themes into unique tokens;
- Extract the average tone score by unique theme token;
- Apply country filter to location field.

Each data set is aggregated to daily frequency.

The GKG's themes field contains a string of unique labels with all themes that the GDELT algorithm identifies for each news item scanned. These themes are very detailed and I leverage their taxonomy to group them into categories such as “government”, “disaster” or “social” (see appendix C.1 for a list of all theme categories). These theme groups represent events or conditions, apart from “actor”, “ethnicity”, “language”, “point of interest” and “animal”, which are purely descriptive and are therefore not included. On average, this cuts down the number of unique themes from around 7,000 to around 5,000 in each country’s data set.

Figure 6.1: Average split of theme categories



Economic and financial themes represent around 35% of all themes on average, followed by themes related to disease and disaster.

6.3.4 Data preprocessing

This section describes the data preprocessing performed.

Applying the augmented Dickey Fuller unit root test, stationarity is rejected at 5% significance for the explanatory variables set out in section 6.3.2 and most of the average tone scores from GDELT. I transform all time series by taking the five business day difference and use the differenced data for modeling BEIR. By doing so, I decrease the noise inherent in daily data and tackle the issue of non-stationarity – implementing the aforementioned test on the differenced time series allows not to reject stationarity for all variables. Then, the variables are standardized by removing the mean and scaling to unit variance.

I look to forecast if a country's respective BEIR will move up or down over the next five business days. Hence, I convert the five day change in each country's BEIR into a binary time series i.e. one (zero) for an increase (decrease) in each BEIR. The two classes are balanced across economies considered, and no case is observed where the BEIR has not moved over five days.

6.4 Analysis approach

This section outlines the analysis that is conducted to investigate if features extracted from GDELT enhance forecasts of BEIR movements.

This is a classification task due to the binary nature of the predicted variable.

I employ five types of machine learning algorithms to forecast changes in BEIR one day ahead (see section 3.5.1 for details):

- Logistic Regression;
- Support Vector classifier;
- Random Forest classifier;
- XGBoost classifier;
- Multilayer Perceptron classifier.

I have explored a range of time intervals, i.e. three, five and ten days and a difference of five days generated the best outcomes across models and countries. Similarly, I have explored a range of prediction windows such as one, three and five days ahead, with one day ahead predictions returning the best results.

When modeling GDELT data, I address the issue of high dimensionality and multicollinearity of features by implementing partial least squares (PLS) as dimensionality reduction method (see section 3.4) to derive components for inclusion in the predictive models set out above.

I derive PLS components from GDELT data in two steps. First, I forecast each respective BEIR using the market-based explanatory variables only and a Logistic Regression model. Second, I apply PLS to the residuals from these predictions as the residuals are the portion of the predicted variable that is unexplained, thus adding new information to the explanatory variables.

The first five PLS components account for over 80% of the variation in each country's target variable. Applying cross-validation analysis shows that the residual sum of squares increases in models with more than five PLS components; therefore five components are appropriate (Tobias, 1995).

For benchmarking purposes, I use each respective framework listed above, only including the market-based explanatory variables.

For performance evaluation, I apply walk-forward cross-validation with a five-fold split (see section 3.6.3) and precision, recall and the F1 scores (see section 3.6.1). I use the McNemar test to gauge whether model predictions are significantly different compared to benchmark predictions (see section 3.6.4).

6.5 Findings

This section contains the findings from the analysis outlined in section 6.4.

In order to predict the five day changes of each country's BEIR, one day ahead, the five PLS components extracted from each respective country GDELT data set and the market-based explanatory variables are incorporated into the predictive frameworks set out in section 6.4.

The column names in Table 6.1 stand for Logistic Regression (LG), Support Vector classifier (SV), Random Forest classifier (RF), XGboost classifier (XG) and Multilayer Perceptron classifier (MLP).

Table 6.1: Results of the models set out in section 6.4.

BEIR for \ Model	LG	SV	RF	XG	MLP
US	0.0947	0.0324	0.1968	0.2158	0.1554
UK	0.1509	0.0337	0.1344	0.1590	0.1743
Germany	0.1464	0.1439	0.2127	0.2097	0.1652
Japan	0.0525	0.0280	0.0717	0.0917	0.0512
South Africa	0.1986	0.1862	0.2578	0.2590	0.2512
Australia	0.0262	0.0341	0.0813	0.0886	0.0690
Brazil	0.2020	0.2383	0.2340	0.2320	0.1935
Mexico	-0.2252	-0.1534	-0.1260	-0.1232	-0.1522

Numbers in blue (red) fields show the improvement (deterioration) in F1 score relative to the respective benchmark.
Features extracted from GDELT themes enhance predictions of the five day change in BEIR, one day ahead, for seven out of eight economies considered. Overall, the Logistic Regression and XGBoost algorithms deliver the best results (see section C.3 in terms of precision, recall and F1 scores for all respective models and benchmarks).

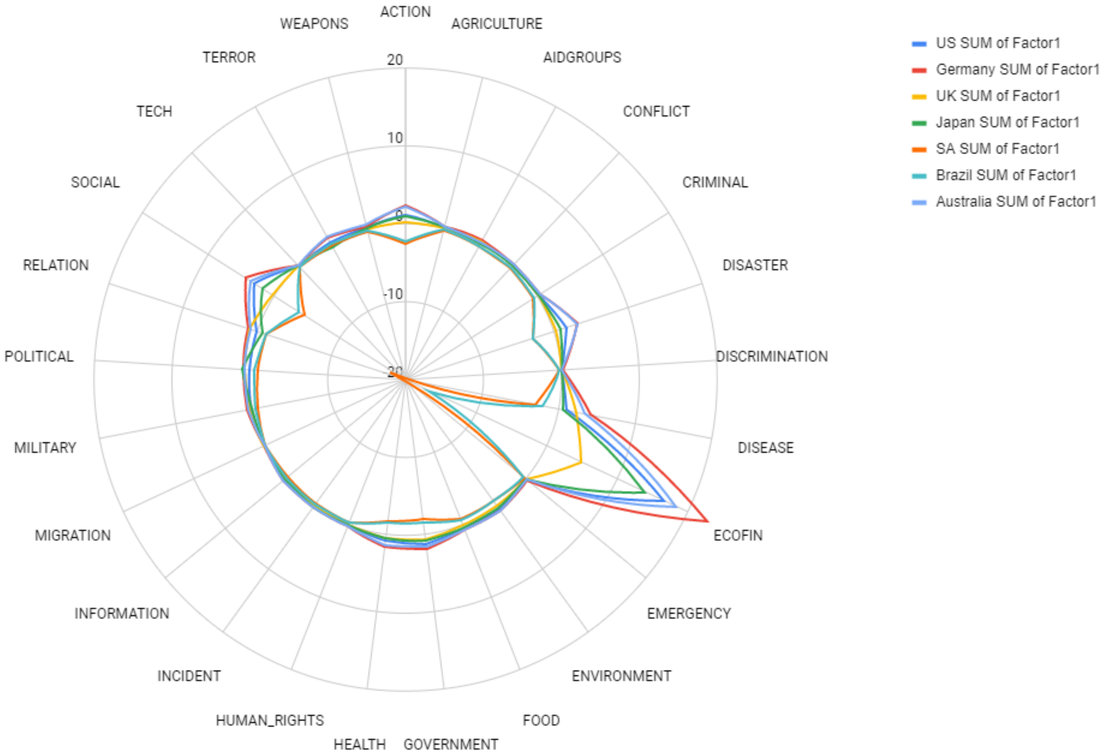
According to the McNemar statistics, model predictions are significantly different to benchmark predictions at a 5% level (see Table C.7).

6.5.1 Feature importance analysis

This section complements my analysis by providing details on the models' performance drivers. For each country, I analyse the coefficients of those models delivering the strongest performance, Logistic Regression and the XGBoost classifier.
In the Logistic Regression models, all PLS components are statistically significant at 1%, with the first PLS component being the largest coefficient in absolute terms, followed by the second and third components.
Likewise, according to each model's feature importances, the first three components are the three most important determinants of performance for XGBoost classifiers. Feature importance measures the increase in accuracy a variable adds to each tree in the classifier, with a higher value indicating a higher importance (Friedman et al., 2001).
In order to gain insights into the relationship between the first PLS component and the original GDELT themes, I examine the loadings. The loadings denote the strength of relationship between each original theme and the components.

GDELT contains c 13,000 very detailed themes, representing events or conditions, which I group into specific categories by the GDELT theme taxonomy. For instance, themes including “health” are grouped into a health category; themes containing the term “weapon” are associated with the weapons category, and so on (see section C.1 for detailed list).

Figure 6.2: Relationship between first PLS component and theme groups



The first PLS component has the strongest relationship with economic and financial (“ecofin”) themes across all countries and theme groups considered.

The radar chart in Fig. 6.2 contains the theme groups linked to the loadings of those first PLS components for those Logistic Regression and XGBoost models that beat their benchmark (see Tables C.1 and C.4).

Fig. 6.2 illustrates that economic and financial themes (“ecofin”) have the strongest connection with the first PLS component across economies. This indicates that economic and financial themes are the main determinant of this component and thus have the strongest predictive capacity, followed by “social”, “health”, and “government” themes. PLS components 2 and 3 are associated with a mix of theme groups, with no particular group being a dominant driver of the respective component (Figures are available on request).

6.5.2 Cross-country impact of narrative on BEIR

In this section, I examine the cross-country spillovers of narrative on BEIR. I utilise Graphical Granger Causality modelling for quantifying this phenomenon (Lozano et al., 2009).

The Granger Causality test is applied to pairs of variables to find out whether there are statistical links between those features (Granger, 1969).

According to Arnold *et al.* (Arnold et al., 2007), I implement the "exhaustive graphical Granger method" on every feature pair using the five day difference in all 74 features and a day's lag. The test's null hypothesis stipulates that a lagged feature does not Granger-cause a feature at 5% significance; the alternative hypothesis stipulates the that a lagged feature Granger-causes a feature at the same level of significance. I apply the Benjamini-Hochberg (BH) procedure to adjust for false positives given multiple tests (Benjamini and Yekutieli, 2005).

If two features have a statistically significant adjusted p-value, the feature pair becomes a directed edge where the first feature is the source node. All features have at least one in- or outbound edge and are therefore kept as nodes.

According to the Graphical Granger Causality modelling described above, I create a directed knowledge graph with 74 nodes and 817 edges (see Fig. 6.3), which is connected with a density of 0.1512. In the following analysis, I refer to the five PLS components derived from GDELT themes and their average tone score for eight economies as narrative-based factors.

To better understand the importance of news narrative-based factors in the graph, I compute the inbetweenness centrality scores for all nodes (Friedman et al., 2001). In terms of this centrality measure, nodes that represent the first three PLS components rank in the top quintile. This backs up the results from section 6.5.1, showing that, across countries, PLS components one, two and three have the best predictive capacity. Those nodes with the lowest inbetweenness centrality scores represent features with low importance in both Logistic Regression and XGBoost models.

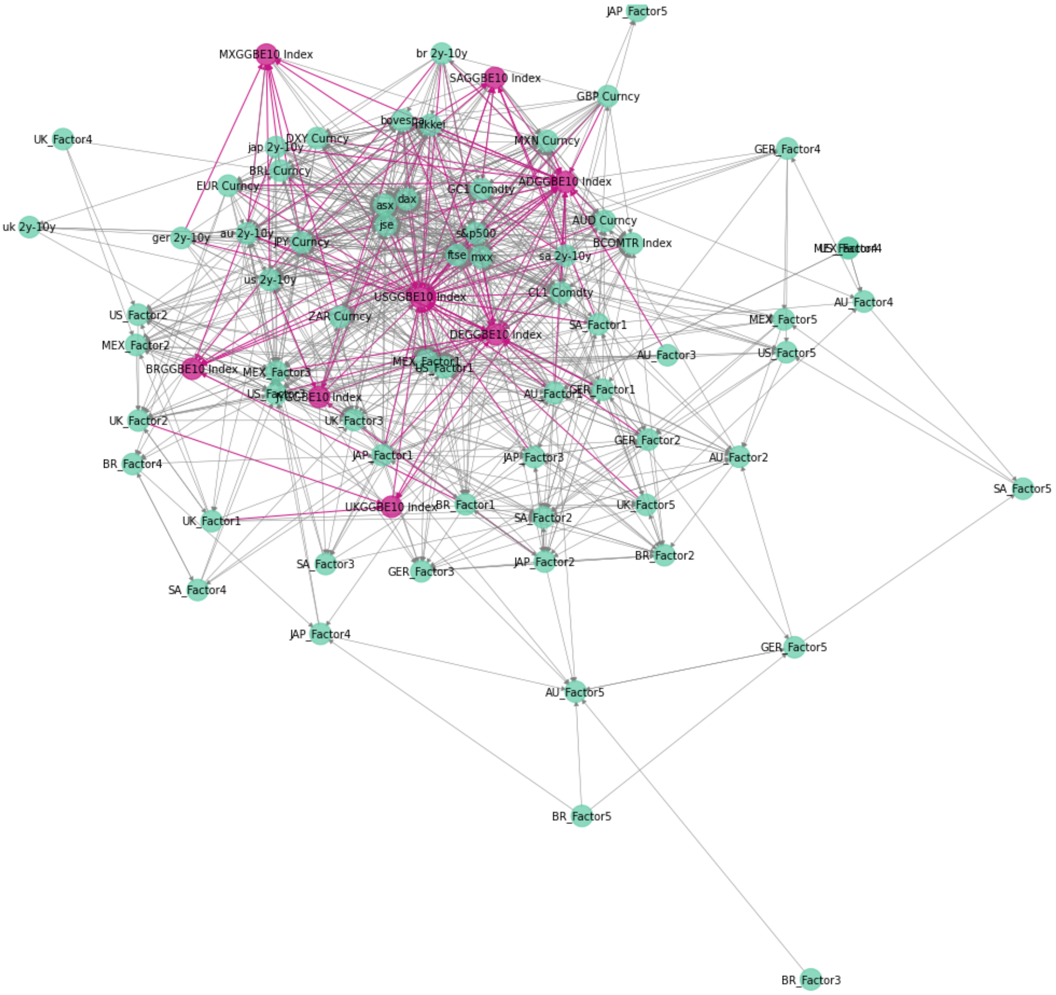
In order to spot cross-country spillover effects of narrative on BEIR movements, I analyse the predecessor nodes for each of the eight BEIR considered. Seven BEIR have incoming edges from market-based explanatory and narrative-based features. Mexico solely has incoming links from market-based features, which is in line with my findings from 6.5 that news narrative does not have a significant effect on Mexico's BEIR.

Only US and German BEIR are impacted by other countries' narrative-based factors as described by incoming edges. For example, the US BEIR is influenced by PLS components from Germany, Japan, South Africa and Australia. The German BEIR is affected by PLS components connected to the US, Japan and Australia. All other BEIR (except of Mexico) have incoming edges from local narrative-based features. The results indicate that GGC modelling helps to identify a cross-country impact from news narrative on BEIR for two out of eight economies considered, the US and Germany.

Following the idea of the global interconnectedness of local capital markets, the incoming edges for all BEIR indicate that cross-country spillover effects may also be explained by market-based features.

The full knowledge graph is shown in Fig. 6.3, with BEIR and their incoming links marked in magenta.

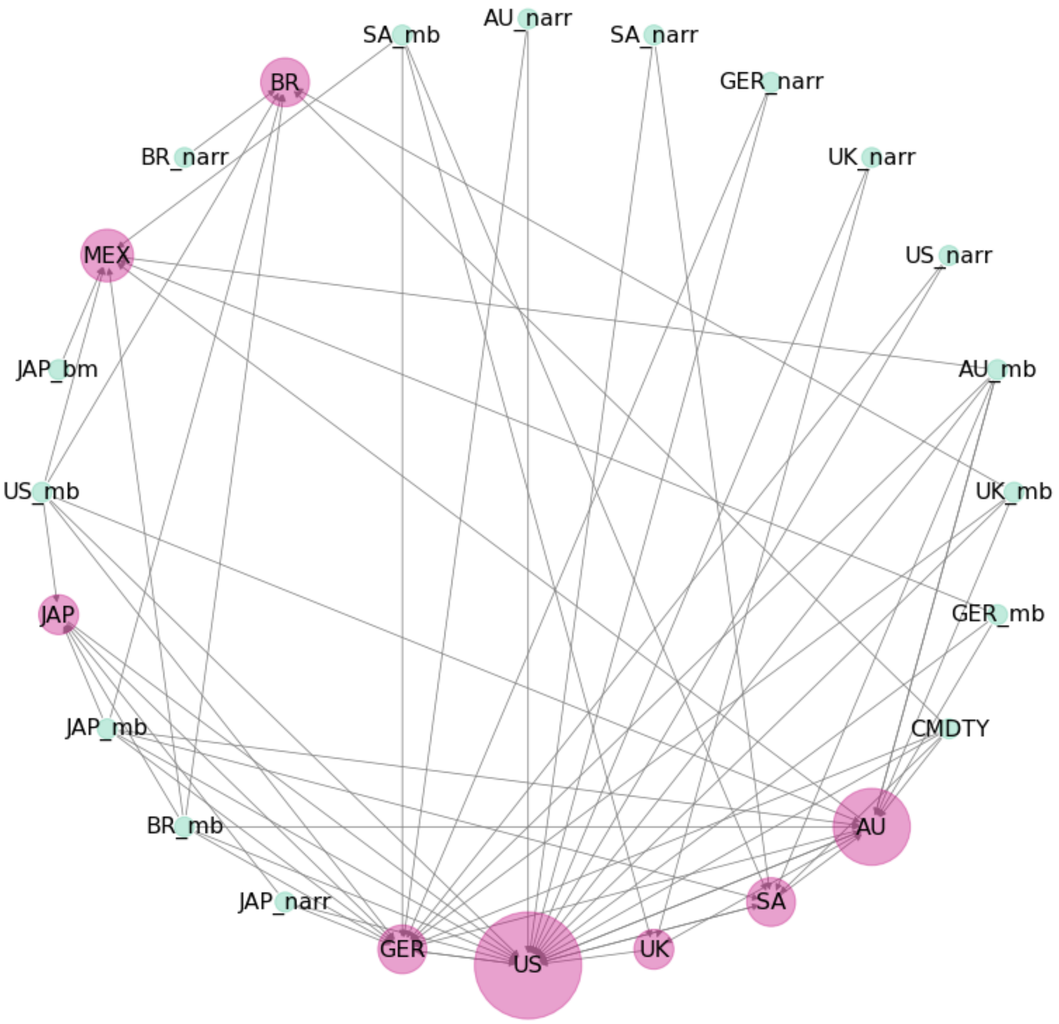
Figure 6.3: Graphical Granger Causality between features



The BEIR and their incoming links marked in magenta show that some of the features Granger-cause BEIR.

Fig. 6.4 contains a simplified version of the knowledge graph, only showing the eight BEIR and their incoming edges. Nodes with country names stand for the respective country's BEIR (magenta nodes), the suffix "_mb" covers all nodes describing country-specific market-based explanatory features, the node including "CMDTY" refer to commodity-related features and the suffix "_narr" denotes country-specific news narratives. The node size represents the amount of incoming edges.

Figure 6.4: Graphical Granger Causality: incoming edges of BEIR



Narrative-related features not related to the US and Germany, respectively, Granger-cause changes in the US and German BEIR. I refer to this as cross-country impact of narrative on BEIR.

My findings emphasise the impact of news narrative on changes in a country's BEIR, with a cross-country impact of news narrative identified for changes in the US and German BEIR. Overall, the results convey that features extracted from news articles can quantify the market's "animal spirits" (Keynes, 1937; Shiller, 2019; Tuckett et al., 2014).

6.6 Discussion

In this experiment, I introduce an effective big data filtering approach to distill large amounts of data and to build data sets reflecting economic and financial news for eight diverse economies. I forecast the five day changes in BEIR for these economies one day ahead using market-based explanatory features and variables based on news narrative, leveraging five different machine learning algorithms. The models generally outperform their respective benchmarks that do not include news-based variables, with Logistic Regression and XGBoost algorithms beating their corresponding benchmarks for seven out of eight economies. I supplement those findings with a feature importance analysis, showing that economic and financial themes are the most important determinants of the first PLS component, which is the key performance driver in Logistic Regression and XGBoost models.

According to these results, I cannot reject H_1 and state that features based on news narrative significantly enhance predictions of changes in market-based inflation expectations.

I leverage GGC modelling to investigate cross-country spillover effects of news narrative on BEIR. My findings suggest that these are present in the case of the US and Germany. Mexico's BEIR is solely affected by market- based features and the other five BEIR are only influenced by local narrative. For this reason, I cannot reject H_2 for the US and Germany but reject it for the other economies covered in this study.

Chapter 7

Generating trading signals with news sentiment

7.1 Introduction, motivation and background

Stock markets are driven by a multitude of factors, whose importance varies over time. While diverse views and beliefs about stock market movements make it impossible to accurately forecast future prices, price moves can be estimated. Theoretical frameworks explain how observed external factors cause prices to deviate from their assumptions. According to the efficient capital market theory, stock markets are very efficient in reflecting information about individual securities and the entire stock market (Malkiel and Fama, 1970).

However, traditional quantitative approaches struggle to explain unexpected variations in prices, and Shiller finds that unsettling narrative precedes events in the economy and markets (Shiller, 2019). This suggests that news narrative and sentiment contain a wealth of information that may not be fully reflected in the stock market yet.

7.2 Literature review

Numerous studies stress the importance of information from news, showing that sentiment derived from news narrative can be useful for predicting stock market returns.

Twitter is a very popular data source for raw narrative and can easily be accessed via API. Nisar and Yeung collect Twitter tweets over six days around the May 2016 local elections in the UK, using three hashtags (trending topics) (Nisar and Yeung, 2018). A sentiment-based lexicon is used to extract positive, negative and neutral sentiment and generate a net mood score. Regression and correlation analysis suggests that there is causation between Twitter sentiment and stock market movements. Elshendy *et al.* extract features from Twitter, Wikipedia, Google Trends and GDELT and attempt to predict the daily crude oil price from April 2013 to April 2015 (Elshendy et al., 2018). Results from a autoregressive framework show that these features improve model performance while being statistically significant. Research by Oliveira *et al.* apply a stock market lexicon to microblogging data from December 2012 to October 2015 and use a range of survey indices (Oliveira et al., 2017). These indices have varying frequency and are merged into one sentiment indicator with a Kalman filter. The study uses the Diebold Mariano test to establish if sentiment-based predictions are useful compared to an autoregression benchmark model. Results show that Twitter sentiment adds value when predicting movements in the broad stock market, smaller market caps as well as sectors. A study by Checkley *et al.* is more critical of the use of social media sentiment for predicting stock price moves, arguing that there is only modest evidence of its predictive power (Checkley et al., 2017). Pagolu *et al.* filter Twitter tweets from August 2015 to August 2016 using keywords (Pagolu et al., 2016). A training set is created, manually annotating tweets as positive, negative or neutral, and the tweets are classified. The sentiment data is used to predict increases or decreases in the Microsoft stock price. The support vector machine algorithm produces the best results with an accuracy of 71.8%. Souza *et al.* find that Twitter sentiment associated with a selection of listed retailers has a statistically significant relationship with the respective stocks' returns and volatility (Souza et al., 2015).

Newspaper archives hosted by data providers feature as another source of raw narrative in publications. Hausler *et al.* classify 12 years of news from the SP Global Markets Intelligence platform into positive, negative and neutral using support vector machines (Hausler et al., 2018). Findings from an autoregressive framework suggest that positive sentiment Granger-causes real estate returns and leads the market by one month.

More recent studies use preprocessed sentiment scores provided by Thomson Reuters to predict stock market moves. Huang *et al.* use the Thomson Reuters Market Psych Index from 1998 to 2016 (Huang et al., 2018). Regressions reveal relationships between sentiment and a range of stock markets and asset classes. Social media sentiment is more persistent for assets held by retail investors than traditional media sentiment. A study by Heston *et al.* maps articles from Thomson Reuters NewsScope (sentiment scores) and Thomson Reuters News Archive together (2003 - 2010) and extracts positive, negative and neutral sentiment (Heston and Sinha, 2017). The data is fed into a neural network to predict excess returns on individual stocks. Weekly aggregated data predicts excess returns up to a quarter, pointing to the detection of under-incorporated information. Positive news affects stock prices within one week, negative news within a quarter. Consoli *et al.* show that negative emotions such as “distress” and “panic” extracted from news help predict movements in government yield bond spread in Italy and Spain, using GDELT as data source (Consoli et al., 2021).

Existing research uses a broad range of data for forecasting financial markets. However, the majority of sources is anglophone, ignoring signals from non-English speaking media. Very few papers use GDELT as a data source.

In this experiment, I attempt to predict a broad range of equity market movements using the average tone score from GDELT, addressing the research question whether the average tone score from GDELT has predictive power for a broad range of equity markets. Accordingly, I state the following hypothesis:

- H_1: Changes in the GDELT average tone score are predictive of changes in equity markets.

Chapter 7 contributes to existing research on predicting equity markets with news sentiment by demonstrating that the average tone score from GDELT is a leading indicator for equity market movements. Based on the GDELT average tone score, I devise a profitable momentum trading strategy for 15 diverse equity markets around the world. Systematic asset managers and hedge funds are continuously searching for new sources of alpha (defined as the excess return of an investment relative to the return of a benchmark) and have recently turned their attention to so-called alternative data sets such as news narrative. GDELT represents a relatively unexplored yet very rich data source and therefore has many applications in this field.

7.3 Data and methods

This section outlines the filtering methodology and the preprocessing applied.

7.3.1 Filtering methodology

This experiment is based on the data obtained by applying the filtering methodology set out in section 4.3.3 to GDELT data, filtering for “economic growth”. The filtered data is then aggregated to daily frequency and country filters are applied, creating average tone time series for the US, the UK, Germany, Spain, Italy, France, the Netherlands, Switzerland, Japan, Hong Kong, Singapore, Taiwan, Australia, South Africa and Canada. The GDELT average tone score is described in detail in section 4.3.4.

7.3.2 Data preprocessing

For all equity markets, the daily percentage change is calculated. For each average tone score, the rolling one-month change is computed. Other rolling time intervals

such as one week, two weeks and three months have been explored and the one-month change in average tone yields the best results.
The augmented Dickey Fuller unit root test is applied to both daily equity market returns and differenced average tone scores and stationarity cannot be rejected for any of the variables at 5% significance.

7.4 Data analysis

This section sets out the analyses conducted to gauge if GDELT average tone scores have a statistical relationship with stock market returns. Further, I assess whether this sentiment indicator can be transformed into a profitable investment strategy.

7.4.1 Causality analysis

The data is tested for Granger Causality using a lag up to three days to examine whether the monthly change in average tone score filtered for one country is predictive of the respective daily equity market returns (see section 3.2). Further, the data is tested for reverse Granger Causality, i.e. a statistical link between a country's daily stock market returns and the respective monthly change in sentiment. The Benjamini-Hochberg (BH) procedure is applied to the generated p-values to adjust for multiple hypothesis testing (see section 3.3). Table 7.1 contains the number of p-values that are statistically significant at a level of 5%.

Table 7.1: Number of p-values significant at 5% level, applying a lag of up to three days

Country	GC	reverse GC
US	3	3
UK	2	3
Germany	3	3
Japan	1	3
Netherlands	1	-
Italy	2	-
Spain	1	-
Hong Kong	-	3
France	3	3
Singapore	-	3
Taiwan	2	3
Switzerland	-	-
Australia	1	3
South Africa	-	-
Canada	3	-

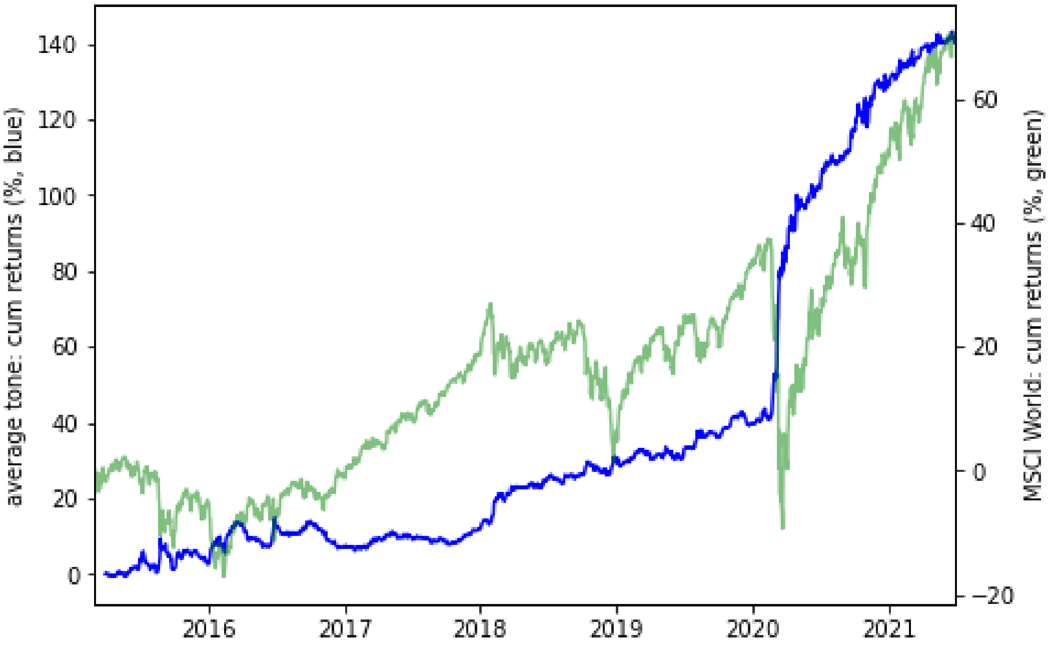
Findings from the Granger Causality tests imply that GDELT's average tone score Granger-causes the respective daily stock market returns for 11 out of 15 countries. Further, the results highlight that reverse Granger Causality exists between daily stock market returns and the one-month change in average tone score for nine out of the 15 countries considered.

7.4.2 Backtest

To transform the GDELT sentiment scores into a trading strategy, I apply a simple rule, i.e. holding a 100% long position in futures of the stock market of the countries mentioned above in case the one-month change in sentiment is positive, otherwise implementing a 100% short position. The trading frequency is weekly with a trading lag of one day. Other time intervals related to the change in sentiment score and different trading frequencies have been explored and the above-mentioned ones yield the best results. Each country is equally weighted and the assumed trading costs are 2bps. Currencies are unhedged.

The backtest goes back to February 2015, which is the inception date of GDELT v2.

Figure 7.1: Backtest performance



The chart shows the cumulative returns of the trading strategy (blue), compared against the cumulative returns of the MSCI World index (green). The strategy performs particularly well during the sell-off of equity markets in March 2020.

According to the backtest, the strategy generates a net annualised return of 14.4% with sharpe ratio of 1.8, a maximum drawdown of 7.5%, a skewness of 2.2, a kurtosis of 17.2 and a turnover of 3,362%.

At a country-level, all countries deliver a positive return, with Spain, Germany and Switzerland being the top and the Netherlands, Taiwan and Italy being the bottom contributors, respectively (see Table D.1 in the appendix for a full performance attribution by country). Figs. D.1 and D.2 in the appendix provide examples of buy and sell signals occurring in individual country strategies, illustrating that they capture the steep sell-off in global equity markets in March 2020 particularly well. The strategy acts as an excellent tail risk hedge during the steep sell-off in equity markets in March 2020 due to the onset of the COVID-19 pandemic, while also being profitable during "normal" periods.

Thus, the sentiment data captures the dynamics during the pandemic well which is impressive as most systematic strategies based on "conventional data sources" struggle during such periods.

7.5 Discussion

In this experiment, I apply the filtering methodology introduced in section 4.3.3 to filter GDELT data for news items relevant to “economic growth”. I then extract the average tone score for 15 diverse countries. The Granger Causality test reveals that a statistical relationship exists between the one-month change in average tone scores and the 15 respective daily stock market returns.

I demonstrate that the average tone scores can be transformed into a profitable trading strategy, which exhibits particularly strong performance during the downturn in global equity markets in March 2020.

Based on my results, I cannot reject H_1 and accept that changes in GDELT’s average tone score are predictive of changes in equity markets.

Chapter 8

General Conclusions

8.1 Summary

The importance and breadth of my research is reflected in the title of this thesis, namely “The impact of news narrative on the economy and financial markets”, which characterises the wide array of topics covered by the experiments in this thesis.

A variety of entities can be extracted from newspaper text, such as emotions, themes and locations. Media reporting is a dynamic mirror of world events and therefore, entities extracted from news narrative capture these dynamics.

The four experiments contained in this thesis provide empirical results to demonstrate the impact of news narrative on the economy and financial markets. The studies are interrelated since they all leverage GDELT data, a very large database that monitors global media and extracts locations, organizations, themes, emotions, events and many other items that drive global society. This thesis consists of a series of exploratory studies and ends by proposing a sentiment-based systematic trading strategy.

Macroeconomic forecasting through news, emotions and narrative

This experiment introduces a novel approach to including emotions extracted from news articles into macroeconomic forecasts, attempting to predict industrial production and consumer prices using narrative and sentiment from global media reporting. The majority of existing works only utilises positive and negative news tone to enhance macroeconomic forecasts and has a distinct focus on the US economy. Sources of narrative are predominantly anglophone and do therefore not capture the global breadth of media reporting. This experiment advances the existing body of literature in this field by including a wide range of sentiment scores extracted from global news narrative by the Global Database of Events, Language and Tone (GDELT) into macroeconomic forecasts. I propose a thematic data filtering approach based on a bi-directional long short term memory neural network (Bi-LSTM) for deriving emotion scores from GDELT and show its effectiveness by contrasting results for filtered and unfiltered data. I forecast industrial production and consumer prices for ten diverse economies using an autoregressive framework, and conclude that incorporating emotions from global newspaper text significantly enhances forecasts compared to three autoregressive benchmark models. I supplement my analysis with an interpretability study on distinct emotion categories, which reveals that emotions linked to happiness and anger have the strongest predictive capacity for the indicators that I forecast.

Macroeconomic forecasting through statistically validated knowledge graphs

This experiment uses narrative from global newspapers to build theme-based knowledge graphs about world events, showing that variables extracted from such graphs improve predictions of industrial production in three large countries compared to several benchmarks. This study leverages a filtering approach that identifies "backbones" of statistically significant edges from large graph data sets. Changes in the eigenvector centrality of nodes in such backbones encapsulate changes in relative importance between different themes significantly better than graph similarity metrics. I complement my findings with an interpretability study, demonstrating that the theme groups “disease” and “economic” have the strongest predictive capacity during the time period considered. This study provides a blueprint for the construction of parsimonious – yet information-rich – theme-based graphs to track the evolution of relevant events in socio-economic systems in real-time.

Predicting market inflation expectations with news topics and sentiment

This experiment proposes a new method of including news themes and their associated sentiment into predictions of changes in breakeven inflation rates (BEIR) for eight diverse economies with mature fixed income markets. I utilise five types of machine learning algorithms incorporating narrative-based features for each economy, and establish that they generally beat their respective benchmarks that do not incorporate such features. Across the countries considered, Logistic Regression and XGBoost classifiers perform best. I conduct a feature importance analysis to add some interpretability to my results, which reveals that economic and financial themes are the key determinants of model performance, with some contributions from themes related to health and government. I analyse the cross-country impact of news reporting on BEIR using Graphical Granger Causality modelling and identify this effect for the US and Germany, while five other economies in my analysis are solely impacted by local news reporting.

Generating trading signals with news sentiment

This experiment proposes a news-sentiment based systematic trading strategy leveraging GDELT data filtered according to the methodology introduced in the first experiment to create a profitable systematic trading strategy based on the average tone scores for 15 diverse economies.

8.2 Future work

The experiments in this thesis propose several possible extensions to address any potential limitations and open questions. First, GDELT version 2 has a limited track record, starting end of February 2015. This relatively short track record very likely impacts the significance of results and does not cover important market events such as the global financial crisis in 2008/9. The database's creators have plans to eventually backfill its history to 1979 and it will be interesting to repeat the experiments once a longer track record has been added. Second, my work only explores data from the GKG. GDELT contains many other information-rich tables such as the event table. The latter in particular could be of high relevance when building entity-based knowledge graphs to monitor change in socio-economic systems and is worthy of further work. Third, the experiments provide proofs of concepts, without fully optimising model performance. Real-world applicability can be improved by including more explanatory variables and expanding to a broader range of predicted variables. Fourth, this thesis uses GDELT as a source of narrative-based features. GDELT is a very rich database that lends itself well to the research on macroeconomic subjects and broad financial markets. However, depending on the nature of the research subject, other sources of narrative may be more appropriate and this deserves further exploration by other researchers – for instance, social media platforms such as Twitter or Reddit seem more suitable for extracting narrative-based features relating to cryptocurrencies.

Appendix A

Forecasting the economy through news, emotions and narrative

A.1 GCAM sentiment scores

This section outlines the GCAM sentiment scores used in this experiment.

ML Senticon

- Level 1 to Level 8 Positive (Spanish)
- Level 1 to Level 8 Negative (Spanish)
- Level 1 to Level 8 Positive (English)
- Level 1 to Level 8 Negative (English)

Hedonometer

- Happiness (English)
- Happiness (French)
- Happiness (German)
- Happiness (Spanish)

- Happiness (Hindu)
- Happiness (Indonesian)
- Happiness (Korean)
- Happiness (Arabic)
- Happiness (Portuguese)
- Happiness (Russian)
- Happiness (Urdu)
- Happiness (Chinese)

Loughran & McDonald Financial Dictionary

- Litigious
- ModalStrong
- ModalWeak
- Negative
- Positive
- Uncertainty

WordNet-Affect

Given the large number of WordNet-Affect scores, I only list some examples for illustrative purposes.

- Abashment
- Abhorrence
- Admiration
- ...

- world-weariness

- worship

- wrath

A.2 *P*-values from modified Diebold Mariano test

Tables A.1 and A.2 contain the *p*-values from the modified Diebold Mariano test. This test is conducted to assert whether model predictions incorporating filtered GDELT sentiment components significantly differ from benchmark predictions set out in section 4.5.2.

Table A.1: *P*-values from modified Diebold Mariano test (IP)

IP for	Model - BM1	Model - BM2	Model - BM3
US	0.0000	0.0003	0.0000
UK	0.1848	0.1007	0.0812
Germany	0.0103	0.0093	0.0091
Norway	0.0178	0.1081	0.0181
Poland	0.0338	0.0094	0.0471
Turkey	0.1014	0.0017	0.1014
Japan	0.0025	0.0812	0.0025
South Korea	0.0450	0.9890	0.0600
Brazil	0.0052	0.0865	0.0053
Mexico	0.0767	0.0472	0.1009

The modified Diebold Mariano test illustrates that model predictions for IP significantly differ from BM1 and BM2 in nine out of ten, and for BM3 in 10 cases, respectively a significance level of 1, 5 or 10%.

Table A.2: *P*-values from modified Diebold Mariano test (CPI)

CPI for	Model - BM1	Model - BM2	Model - BM3
US	0.0132	0.4314	0.9216
UK	0.0238	0.1730	0.0349
Germany	0.0701	0.1014	0.0013
Norway	0.0111	0.0599	0.0091
Poland	0.0072	0.0789	0.0027
Turkey	0.0169	0.0017	0.0010
Japan	0.0152	0.0812	0.0342
South Korea	0.0161	0.0345	0.2579
Brazil	0.0029	0.0481	0.0599
Mexico	0.0028	0.8028	0.9100

Model predictions for CPI differ from BM1 for all countries and for BM2 and BM3 for seven out of ten countries, respectively at a significance level of 1, 5 or 10%.

Appendix B

Macroeconomic forecasting with statistically validated knowledge graphs

B.1 COVID-19 related themes

The themes listed below represent symptoms associated with COVID-19. Using these themes, I calculate the median eigenvector centralities for each calendar month's graph and, on the case of the COVID-19 outbreak in 2020, show the evolution of these centrality metrics over time.

- `WB_2165_HEALTH_EMERGENCIES`
- `WB_1406_DISEASES`
- `TAX_DISEASE_CORONAVIRUS`
- `TAX_DISEASE_EPIDEMIC`
- `TAX_DISEASE_OUTBREAK`
- `TAX_DISEASE_INFECTION`
- `TAX_DISEASE_PNEUMONIA`

- TAX_DISEASE_FEVER
- TAX_DISEASE_INFECTIOUS
- TAX_DISEASE_FLU
- TAX_DISEASE_COUGH

B.2 GDELT theme categories

In this experiment, GDELT themes are associated with 26 theme groups. For instance, the “Weapons” group includes 81 themes like "TAX_WEAPONS_GUNS", "TAX_WEAPONS_BOMB" and "TAX_WEAPONS_SUICIDE_BOMB". For my analysis, I removed purely descriptive themes related to “Actor”, “Language”, “Animal” and “Ethnicity” groups.

- Economic
- Disease
- Actor
- Language
- Ethnicity
- Animal
- Disaster
- Social
- Relation
- Political
- Health
- Weapons

- Military
- Terror
- Environment
- Food
- Government
- Aid groups
- Information
- Conflict
- Emergency
- Human rights
- Migration
- Legal
- Criminal
- Other (various events or conditions, c 6% of themes)

B.3 *P*-values from modified Diebold Mariano test

Table B.1 shows the *p*-values from the modified Diebold Mariano test, which is applied to establish if factor enhanced model predictions are significantly different from the benchmark forecasts as outlined in section 5.5.

Table B.1: *P*-values from modified Diebold Mariano test

IP for	Model - BM1	Model - BM2	Model - BM3
US	0.0000	0.0000	0.0000
Germany	0.0103	0.0093	0.0001
Japan	0.0000	0.0000	0.0887

Model predictions for IP differ for all benchmarks and all three countries considered at a significance level of 1% or 5%.

Appendix C

Predicting market inflation expectations with news topics and sentiment

C.1 GDELT theme categories

In this experiment, GDELT themes are categorised into 30 theme groups. The difference in the number of theme categories compared to the previous experiment is due to the different filter that is applied to GDELT data and to the fact that themes previously in “other“ are now assigned to theme groups. Purely descriptive themes of categories “Actor”, “Language”, “Animal” “Points of Interest” and “Ethnicity” have been omitted. Theme groups are:

- Ecofin
- Disease
- Actor
- Action
- Language
- Ethnicity

- Animal
- Disaster
- Social
- Relation
- Political
- Health
- Weapons
- Military
- Terror
- Environment
- Food
- Government
- Aid groups
- Information
- Conflict
- Emergency
- Human rights
- Migration
- Agriculture
- Discrimination
- Incident

- Criminal
- Tech
- Points of interest

C.2 Model specifications

This section contains details on the algorithms used for predicting movements in BEIR. Classes are balanced.

C.2.1 Logistic Regression

The Logistic Regression model uses the L_2 penalty and the limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfg) solver (Fletcher, 2013).

C.2.2 Support Vector classifier

In this model, I set C at 1.0 and use a radial basis function as kernel.

C.2.3 Random Forest classifier

I employ a bootstrapped Random Forest classifier with 50 trees and a minimum sample split of 2.

C.2.4 XGBoost classifier

I use an XGBoost classifier with a maximum tree depth of 10 for base learners.

C.2.5 Multilayer Perceptron classifier

I employ a Multilayer Perceptron classifier that optimizes the log-loss function using lbfgs (Fletcher, 2013). I utilise 10 hidden layers, relu as activation function, adam as solver and an L_2 penalty of 10^{-4}.

C.3 F1, Recall and Precision scores

The column names in the following tables stand for Logistic Regression, Support Vector classifier, Random Forest classifier, XGBoost classifier and Multilayer Perceptron classifier, respectively.

Table C.1: F1 scores for the models set out in section 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.8124	0.7427	0.8215	0.8400	0.8160
UK	0.7174	0.5726	0.6552	0.6800	0.7166
Germany	0.7096	0.6567	0.7347	0.7226	0.7023
Japan	0.5582	0.5455	0.5417	0.5644	0.5793
South Africa	0.8300	0.8289	0.8704	0.8723	0.8741
Australia	0.6533	0.6569	0.6492	0.6512	0.6629
Brazil	0.6627	0.5403	0.7074	0.6947	0.6676
Mexico	0.3835	0.4483	0.4196	0.4216	0.4074

Across countries, logistic regression and XGBoost models exhibit the strongest performance.

Table C.2: Recall scores for the models outlined in 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.8064	0.7551	0.7829	0.8138	0.8049
UK	0.7047	0.5568	0.6450	0.6786	0.7057
Germany	0.7016	0.6386	0.7474	0.7412	0.7060
Japan	0.5537	0.5260	0.5409	0.5761	0.5932
South Africa	0.8855	0.8184	0.8727	0.8771	0.8872
Australia	0.6468	0.6513	0.6523	0.6475	0.6564
Brazil	0.7122	0.5257	0.7465	0.7566	0.7432
Mexico	0.3846	0.4600	0.4306	0.4568	0.4111

Table C.3: Precision scores for the models described in 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.8246	0.7383	0.8721	0.8731	0.8330
UK	0.7328	0.5920	0.6747	0.6851	0.7301
Germany	0.7250	0.6884	0.7297	0.7130	0.7069
Japan	0.5751	0.5834	0.5589	0.5678	0.5797
South Africa	0.8746	0.8480	0.8687	0.8686	0.8617
Australia	0.6682	0.6672	0.6501	0.6600	0.6742
Brazil	0.6208	0.5732	0.6741	0.6454	0.6080
Mexico	0.3837	0.4408	0.4101	0.4084	0.4049

Table C.4: F1 scores for the benchmark models outlined in 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.7177	0.7103	0.6247	0.6242	0.6606
UK	0.5765	0.5389	0.5208	0.5210	0.5423
Germany	0.5632	0.5128	0.5230	0.5129	0.5371
Japan	0.5057	0.5175	0.4700	0.4727	0.5281
South Africa	0.6314	0.6427	0.6126	0.6133	0.6229
Australia	0.6271	0.6228	0.5679	0.5626	0.5939
Brazil	0.4607	0.3020	0.4734	0.4627	0.4741
Mexico	0.6087	0.6017	0.5456	0.5448	0.5596

Across countries, the logistic regression model exhibits the strongest performance.

Table C.5: Recall scores for the benchmark models explained in 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.6685	0.8237	0.6422	0.6411	0.7156
UK	0.5923	0.5057	0.4931	0.5090	0.5085
Germany	0.6188	0.4383	0.5137	0.5039	0.5074
Japan	0.4925	0.4987	0.4674	0.4718	0.5188
South Africa	0.6549	0.6145	0.5992	0.6177	0.5986
Australia	0.6411	0.6542	0.5729	0.5720	0.6123
Brazil	0.5507	0.2282	0.4751	0.4703	0.4369
Mexico	0.6000	0.5873	0.5399	0.5414	0.5257

Table C.6: Precision scores for the benchmark models set out in 6.4.

Model BEIR for	LG	SV	RF	XG	MLP
US	0.7847	0.6320	0.6171	0.6159	0.6295
UK	0.5631	0.5837	0.5556	0.5370	0.5855
Germany	0.5213	0.6339	0.5416	0.5283	0.5812
Japan	0.5223	0.5409	0.4750	0.4784	0.5401
South Africa	0.6180	0.6803	0.6363	0.6232	0.6661
Australia	0.6243	0.6000	0.5711	0.5624	0.5827
Brazil	0.4298	0.5823	0.4767	0.4608	0.5294
Mexico	0.6213	0.6219	0.5564	0.5528	0.6058

C.4 McNemar statistics

Table C.7 shows the McNemar statistics for all models, conveying that all classifier predictions statistically differ from the benchmark predictions at a significance level of 5% or 1%.

Table C.7: McNemar statistics

Model BEIR for	LG	SV	RF	XG	MLP
US	0.0264	0.0253	0.0328	0.0322	0.0303
UK	0.0236	0.0246	0.0191	0.0213	0.0242
Germany	0.0192	0.0222	0.0266	0.0252	0.0218
Japan	0.0300	0.0295	0.0309	0.0303	0.0282
South Africa	0.0356	0.0367	0.0386	0.0355	0.0352
Australia	0.0107	0.0158	0.0160	0.0167	0.0156
Brazil	0.0149	0.0208	0.0215	0.0243	0.0175
Mexico	0.0231	0.0235	0.0203	0.0194	0.0226

The columns names stand for Logistic Regression, Support Vector classifier, Random Forest classifier, XGBoost classifier and Multilayer Perceptron classifier, respectively.
Numbers show McNemar statistics, comparing model to benchmark predictions. All model predictions are statistically different to benchmark predictions at a 1% or 5% significance level.

Appendix D

Generating trading signals with news sentiment

D.1 Country level performance attribution

Table D.1 outlines each country's contribution to overall returns, assuming an equal country weights.

Table D.1: Performance attribution by country

Country	Return (%)
Spain	1.73
Germany	1.64
Switzerland	1.59
France	1.54
Hong Kong	1.44
Japan	1.36
South Africa	0.97
Canada	0.89
US	0.79
Singapore	0.78
Australia	0.65
UK	0.62
Italy	0.56
Taiwan	0.54
Netherlands	0.02
Sum	15.12

Numbers show each country's contribution to the strategy's gross return in percentage terms. Top performance contributors are Spain and Germany, while the Netherlands contributes the least.

D.2 Examples of trading signals in country strategies

This section illustrates the occurrence of trading signals in individual country strategies, on the examples of the two best performing countries.

Figure D.1: Backtest performance - Spain

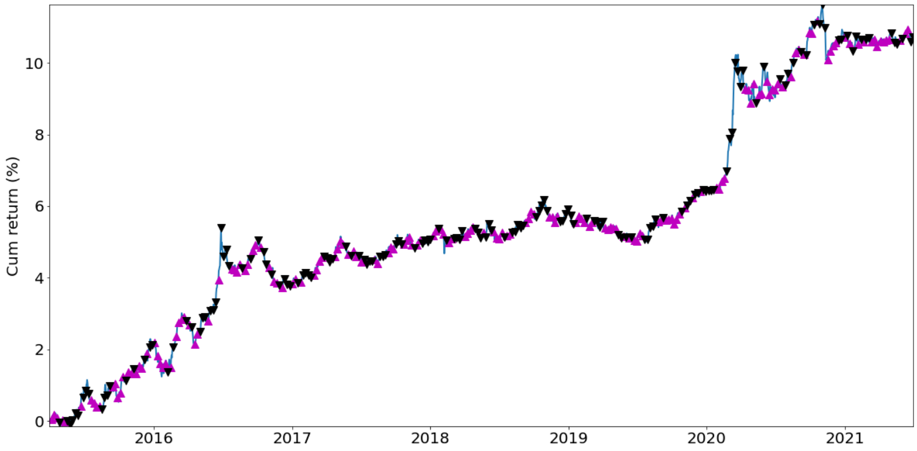


Fig D.1 illustrates the performance of the strategy applied to the main Spanish stock market index, IBEX. The red (black) triangles denote buy (sell) signals.

Figure D.2: Backtest performance - Germany

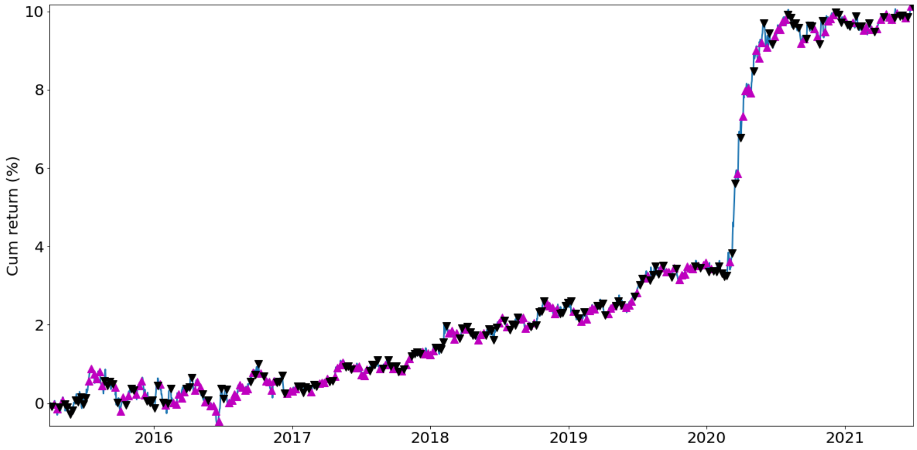


Fig D.2 illustrates the performance of the strategy applied to the main German stock market index, DAX. The red (black) triangles denote buy (sell) signals.

References

Adamic, L., Brunetti, C., Harris, J. H., and Kirilenko, A. (2017). Trading networks. *The Econometrics Journal*, 20(3):S126–S149.

Akerlof, G. A. and Shiller, R. J. (2010). *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Princeton University Press.

Aliber, R. Z. and Kindleberger, C. P. (2017). *Manias, panics, and crashes: A history of financial crises*. Springer.

Allen, D. E., McAleer, M., and Singh, A. K. (2019). Daily market news sentiment and stock prices. *Applied Economics*, 51(30):3212–3235.

Ardia, D., Bluteau, K., and Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, 35(4):1370–1386.

Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75.

Bagrow, J. P. and Bollt, E. M. (2019). An information-theoretic, all-scales approach to comparing networks. *Applied Network Science*, 4(1):1–15.

Baker, S., Bloom, N., Davis, S., and Terry, S. (2020). Covid-induced economic uncertainty and its consequences. https://voxeu.org/article/covid-induced-economic-uncertainty-and-its-consequences.

Baker, S., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701.

Bauer, M. (2015). Inflation expectations and the news. *International Journal of Central Banking*, 39.

Baum, C. (2018). *KPSS: Stata module to compute Kwiatkowski-Phillips-Schmidt-Shin test for stationarity*. Boston College Department of Economics.

Bellomarini, L., Benedetti, M., Gentili, A., Laurendi, R., Magnanimi, D., Muci, A., and Sallinger, E. (2020). Covid-19 and company knowledge graphs: Assessing golden powers and economic impact of selective lockdown via ai reasoning. *arXiv preprint arXiv:2004.10119*.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Bildirici, M. E., Kayıkçı, F., and Onat, I. Ş. (2015). Baltic dry index as a major economic policy indicator: the relationship with economic growth. *Procedia-Social and Behavioral Sciences*, 210:416–424.

BIS (2011). Bis quarterly review. https://www.bis.org/publ/qtrpdf/r_qt1103a.pdf. Accessed 27/05/2021.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170.

Bonaccorsi, G., Riccaboni, M., Fagiolo, G., and Santoni, G. (2019). Country centrality in the international multiplex network. *Applied Network Science*, 4(1):1–42.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Brooks, C. and Tsolacos, S. (2010). *Real Estate Modelling and Forecasting*. Cambridge University Press.

Brosch, T., Scherer, K. R., Grandjean, D. M., and Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss Medical Weekly*, 143.

Bruner, J. S. (1990). *Acts of meaning*, volume 3. Harvard University Press.

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G. L., and Papailias, F. (2018). Evaluation of nowcasting/flash estimation based on a big set of indicators. 16th Conference of IAOS.

Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., and Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1(2017):93–145.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2020). The structure of economic news. https://www.nber.org/papers/w26648.

Campi, M., Dueñas, M., and Fagiolo, G. (2020). How do countries specialize in agricultural production? A complex network analysis of the global agricultural product space. *Environmental Research Letters*, 15(12).

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, 118(1):269–298.

Carvalho, V. M., Nirei, M., Saito, Y. U., and Tahbaz-Salehi, A. (2021). Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics*, 136(2):1255–1321.

Casanova, C., Ortiz, A., Rodrigo, T., Xia, L., and Iglesias, J. (2017). Tracking chinese vulnerability in real time using big data. https://www.bbvaresearch.com/wp-content/uploads/2017/10/Tracking-Chinese-Vulnerability-in-Real-Time-Using-Big-Data.pdf.

Checkley, M. S., Añón Higón, D., and Alles, H. (2017). The hasty wisdom of the mob: How market sentiment predicts stock market behavior. *Expert Systems With Applications*, 77:256–263.

Chen, H.-Y. and Lo, T.-C. (2019). Online search activities and investor attention on financial markets. *Asia Pacific Management Review*, 24(1):21–26.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Ciccarelli, M. and Garcia, J. A. (2009). What drives euro area break-even inflation rates? https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp996.pdf.

Claeskens, G., Hjort, N. L., et al. (2008). Model selection and model averaging. *Cambridge Books*.

Clore, G. L. and Palmer, J. (2009). Affective guidance of intelligent agents: How emotion controls cognition. *Cognitive Systems Research*, 10(1):21–30.

Colladon, A. F., Guardabascio, B., and Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems*, 123:113075.

Consoli, S., Pezzoli, L. T., and Tosetti, E. (2021). Emotions in macroeconomic news and their impact on the European bond market. *Journal of International Money and Finance*, 118.

Constantin, A., Peltonen, T. A., and Sarlin, P. (2018). Network linkages to predict bank distress. *Journal of Financial Stability*, 35:226–241.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Cubadda, G. and Guardabascio, B. (2012). A medium-n approach to macroeconomic forecasting. *Economic Modelling*, 29(4):1099–1105.

da Cunha Cabral, I., Ribeiro, P. P., and Nicolau, J. (2021). Changes in inflation compensation and oil prices: short-term and long-term dynamics. *Empirical Economics*, pages 1–23.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346):1413–1420.

Datawheel, Simoes, A., and Hidalgo, C. A. (2012). The observatory of economic complexity. https://oec.world. Accessed on 15/09/2020.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Doms, M. E. and Morin, N. J. (2004). Consumer sentiment, the economy, and the news media. *FRB of San Francisco Working Paper*, (9).

Doornik, J. A. and Hendry, D. F. (2015). Statistical model selection with "big data". *Cogent Economics & Finance*, 3(1).

Dräger, L. and Lamla, M. J. (2017). Imperfect information and consumer inflation expectations: Evidence from microdata. *Oxford Bulletin of Economics and Statistics*, 79(6):933–968.

Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359.

D'Acunto, F., Malmendier, U., Ospina, J., and Weber, M. (2019). Exposure to daily price changes and inflation expectations. https://www.nber.org/system/files/working_papers/w26237/w26237.pdf.

Ekman, P. and Corduro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370.

Elshendy, M., Colladon, A. F., Battistoni, E., and Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, 44(3):408–421.

Elshendy, M. and Fronzetti Colladon, A. (2017). Big data analysis of economic news: Hints to forecast macroeconomic indicators. *International Journal of Engineering Business Management*, 9:1–12.

Emmert-Streib, F., Tripathi, S., Yli-Harja, O., and Dehmer, M. (2018). Understanding the world economy in terms of networks: A survey of data-based network science approaches on economic networks. *Frontiers in Applied Mathematics and Statistics*, 4:37.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Fraiberger, S. P., Lee, D., Puy, D., and Ranciere, R. (2018). Media sentiment and international asset prices. https://www.imf.org/en/Publications/WP/Issues/2018/12/10/Media-Sentiment-and-International-Asset-Prices-46454.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Fronzetti Colladon, A., Grassi, S., Ravazzolo, F., and Violante, F. (2020). Forecasting financial markets with semantic network analysis in the covid—19 crisis. *https://arxiv.org/abs/2009.04975*.

Gennaioli, N. and Shleifer, A. (2018). *A crisis of beliefs*. Princeton University Press.

Ghalmane, Z., Cherifi, C., Cherifi, H., and El Hassouni, M. (2020). Extracting backbones in weighted modular complex networks. *Scientific Reports*, 10(1):1–18.

Girardi, A., Guardabascio, B., and Ventura, M. (2016). Factor-augmented bridge models (FABM) and soft indicators to forecast italian industrial production. *Journal of Forecasting*, 35(6):542–552.

Glaeser, E. L., Kim, H., and Luca, M. (2017). Nowcasting the local economy: Using yelp data to measure economic activity. https://www.nber.org/papers/w24010.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*, volume 4, pages 2047–2052. IEEE.

Guo, L. and Vargo, C. J. (2020). Predictors of international news flow: Exploring a networked global media system. *Journal of Broadcasting & Electronic Media*, 64(3):418–437.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Hausler, J., Ruscheinsky, J., and Lang, M. (2018). News-based sentiment analysis in real estate: a machine learning approach. *Journal of Property Research*, 35(4):344–371.

He, H. and Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Heston, S. L. and Sinha, N. R. (2017). News versus sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73:3:67–83.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M., and Nelles, O. (2019). Benchmarking in classification and regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1318.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Huang, D., Lehkonen, H., Pukthuanthong, K., and Zhou, G. (2018). Sentiment across asset markets. http://dx.doi.org/10.2139/ssrn.3185140.

Ismail, A. A., Wood, T., and Bravo, H. C. (2018). Improving long-horizon forecasts with expectation-biased lstm networks. *arXiv preprint arXiv:1804.06776*.

Jansen, D.-J. and Neuenkirch, M. (2017). News consumption, political preferences, and accurate views on inflation. *Universitaet Trier, Research Papers in Economics*, 17(3).

Kabiri, A., James, H., Landon-Lane, J., Tuckett, D., and Nyman, R. (2020). The role of sentiment in the economy: 1920 to 1934. https://www.cesifo.org/en/publikationen/2020/working-paper/role-sentiment-economy-1920-1934.

Kapetanios, G. and Papailias, F. (2018). Big data & macroeconomic nowcasting: Methodological review. *Economic Statistics Centre of Excellence (ESCoE)*, 12.

Keynes, J. M. (1937). The general theory of employment. *The Quarterly Journal of Economics*, 51(2):209–223.

Lamla, M. J. and Lein, S. M. (2014). The role of media for consumers' inflation expectation formation. *Journal of Economic Behavior & Organization*, 106:62–77.

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

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

Leetaru, K. H. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. *D-lib Magazine*, 18(9):5.

Leetaru, K. H. (2015a). Gdelt 2.0 Global Knowledge Graph Codebook. https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime. Accessed 15/02/2020.

Leetaru, K. H. (2015b). Gdelt project. https://www.gdeltproject.org. Accessed 15/05/2020.

Leetaru, K. H. (2015c). Mining libraries: Lessons learned from 20 years of massive computing on the world's information. *Information Services & Use*, 35(1-2):31–50.

Leetaru, K. H. (2016). *Can we forecast conflict? A framework for forecasting global human societal behavior using latent narrative indicators*. PhD thesis, University of Illinois at Urbana-Champaign.

Leetaru, K. H., Perkins, T., and Rewerts, C. (2014). Cultural computing at literature scale: encoding the cultural knowledge of tens of billions of words of academic literature. *D-lib Magazine*, 20(9):8.

Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., and Roberts, S. (2014). Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*, 3(2):109–115.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):110–118.

Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, pages 1413–1418.

Maitra, S. and Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, 79:79–90.

Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Mankiw, G. and Reis, R. (2003). *Sticky information: a model of monetary nonneutrality and structural slumps", Knowledge, Information, and Expectations in Modern Macroeconomics: Essays in Honor of Edmund S. Phelps*. Princeton University Press.

Mariano, R. S. (2002). Testing forecast accuracy. *A companion to economic forecasting*, 2:284–298.

Matesanz Gomez, D., Ferrari, H. J., Torgler, B., and Ortega, G. J. (2017). Synchronization and diversity in business cycles: a network analysis of the european union. *Applied Economics*, 49(10):972–986.

Mazumder, S. (2021). The reaction of inflation forecasts to news about the Fed. *Economic Modelling*, 94:256–264.

Mihailov, A., Rumler, F., and Scharler, J. (2011). The small open-economy new Keynesian Phillips curve: Empirical evidence and implied inflation dynamics. *Open Economies Review*, 22(2):317–337.

MSCI (2021). Msci market classification. https://www.msci.com/market-classification/Accessed 06/01/2021.

Nisar, T. M. and Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, 4:101–119.

Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127.

Nyman, R. and Ormerod, P. (2020). Text as data: a machine learning-based approach to measuring uncertainty. *arXiv preprint arXiv:2006.06457*.

Oliveira, N., Cortez, P., and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems With Applications*, 73:125–144.

Oughali, M. S., Bahloul, M., and El Rahman, S. A. (2019). Analysis of nba players and shot prediction using random forest and xgboost models. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5. IEEE.

Pagolu, V. S., Challa, K. N. R., Panda, G., and Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements. *International conference on Signal Processing, Communication, Power and Embedded System (SCOPES).*

Pekar, V. and Binner, J. (2017). Forecasting consumer spending from purchase intentions expressed on social media. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–101. Association for Computational Linguistics.

Pfajfar, D. and Santoro, E. (2013). News on inflation and the epidemiology of inflation expectations. *Journal of Money, Credit and Banking*, 45(6):1045–1067.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Piccardi, C. and Tajoli, L. (2018). Complexity, centralization, and fragility in economic networks. *PloS one*, 13(11).

Pimco (2021). Pimco market intelligence: Understanding inflation-linked bonds. https://www.pimco.co.uk/en-gb/resources/education/understanding-inflation-linked-bonds//Accessed 26/05/2021.

Reinhart, C. M. and Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*. Princeton University Press.

Rousidis, D., Koukaras, P., and Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9):6279–6311.

Salisu, A. A., Isah, K. O., Oyewole, O. J., and Akanni, L. O. (2017). Modelling oil price-inflation nexus: The role of asymmetries. *Energy*, 125:97–106.

Schaer, O., Kourentzes, N., and Fildes, R. (2019). Demand forecasting with user-generated online information. *International Journal of Forecasting*, 35(1):197–212.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85(2):461.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Serrano, M. Á., Boguná, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.

Shiller, R. J. (2019). *Narrative economics: How stories go viral and drive major economic events*. Princeton University Press.

Shiller, R. J. and Beltratti, A. E. (1992). Stock prices and bond yields: Can their comovements be explained in terms of present value models? *Journal of Monetary Economics*, 30(1):25–46.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Slaper, T., Bianco, A., and Lenz, P. (2018). Digital vapor trails: Using website behavior to nowcast entrepreneurial activity. In *2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018)*, pages 107–113. Editorial Universitat Politècnica de València.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606.

Souza, T. T. P., Kolchyna, O., Treleaven, P. C., and Aste, T. (2015). Twitter sentiment analysis applied to finance: A case study in the retail industry. *arXiv preprint arXiv:1507.00784*.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Strapparava, C. and Valitutti, A. (2004). Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer.

Tantardini, M., Ieva, F., Tajoli, L., and Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, 9.

Temizsoy, A., Iori, G., and Montes-Rojas, G. (2017). Network centrality and funding rates in the e-mid interbank market. *Journal of Financial Stability*, 33:346–365.

Thorsrud, L. A. (2016). Nowcasting using news topics. big data versus big bank. *Norges Bank Working Paper*, (20/2016).

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proceedings of the 20th Annual SAS Users Group International Conference*, volume 20. SAS Institute Inc Cary.

Tuckett, D., Ormerod, P., Smith, R., and Nyman, R. (2014). Bringing social-psychological variables into economic modelling: Uncertainty, animal spirits and the recovery from the great recession. *Economic Growth eJournal*.

Tumminello, M., Micciche, S., Lillo, F., Piilo, J., and Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, 6(3).

UNECE (2021). United Nations Economic Commission for Europe big data taxonomy. https://statswiki.unece.org/display/bigdata. Accessed 15/06/2021.

Van Eyden, R., Difeto, M., Gupta, R., and Wohar, M. E. (2019). Oil price volatility and economic growth: Evidence from advanced economies using more than a century's data. *Applied Energy*, 233:612–621.

Wiener, N. (1956). The theory of prediction. *Modern mathematics for engineers*.

Witten, I. H., Frank, E., Hall, M. A., Pal, C., and DATA, M. (2005). *Practical machine learning tools and techniques*, volume 2.

Woodford, M. (2001). Imperfect common knowledge and the effects of monetary policy. https://www.nber.org/papers/w8673.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yang, Y., Pang, Y., and Huang, G. (2020). The knowledge graph for macroeconomic analysis with alternative big data. *arXiv preprint arXiv:2010.05172*.