News, Sentiment and the Real Economy

by

Hanjo Odendaal

Dissertation presented for the degree of Doctor of Philosophy in Economics in the Faculty of Economics at Stellenbosch University

Supervisor:	Prof. Johann F. Kirsten
Co-supervisor:	Prof. Monique Reid

December 2020

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2020-07-12 . . . .

Abstract

News, Sentiment and the Real Economy

H. Odendaal
Department of Economics,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Dissertation: PhD

December 2020

In this dissertation, text analysis is presented as a complement to traditional survey-based methods used to capture sentiment. This is achieved by firstly constructing media-based sentiment indices from a large variety of news sources for South Africa and presenting these indices as a feasible way to replicate the results of the traditional survey-based consumer confidence index (CCI). The findings of the cointegration and Granger-causality tests support the hypothesis that news-based indices could possibly be used to address shortcomings commonly experienced in the survey-based alternative. The second contribution towards the literature is the evaluation of the adequacy of media-based indices as a predictor of personal consumption. The predictive power of media sentiment indices are evaluated in a Bayesian forecasting horse race alongside the CCI. The conclusion revealed that the inclusion of media-based sentiment indices as predictors in a model can decrease forecasting errors of personal consumption expenditure. The forecasting errors decreased in the cases of both short and long (up to 2 years) forecasting horizons. The results substantiate the theory that news media sentiment contains information on the coincidental and future state of the economy which is not captured in the CCI. The results suggest that media based indices could function as both a complement or alternative to the CCI in consumption forecasting. The final contribution of the thesis showcases the effectiveness of utilizing domain-specific dictionaries to capture sentiment. In the last chapter, domain-specific dictionaries are constructed, in an automated fashion, using Random Forests. These domain-specific indices successfully capture economic sentiment more accurately than the widely used Loughran dictionary. This framework reduces the resources required to extract information from media reports into a sentiment dictionary, while also maintaining a level of transparency.

Collectively, the results presented in this dissertation offer some initial support for the use of text analysis in South Africa as an alternative way of capturing softer economic indicators such as economic sentiment.

Uittreksel

Nuus, Sentiment en die Reele Ekonomie

(News, Sentiment and the Real Economy

H. Odendaal
Departement Ekonomie,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Proefskrif: PhD

Desember 2020

Hierdie tesis het dit ten doel om te bewys dat teksanalise ’n aanvullende rol kan bied vir tradisionele opname-gebaseerde metodes, as ’n manier om sentiment vas te lê. Dit word gedoen deur media-gebaseerde sentiment-indekse uit ’n groot verskeidenheid nuusbronne in Suid-Afrika op te stel en hierdie indekse voor te stel as uitvoerbare duplikate van die tradisionele indeks vir verbruikersvertroue indeks (VVI). Die bevindinge van die co-integrasie en Granger-causality toetse ondersteun die hipotese dat nuusgebaseerde indekse gebruik kan word om die tekortkominge in die huidige opname aan te spreek. Die tweede bydrae tot die literatuur is die evaluering van die uitvoerbaarheid van mediagebaseerde indekse, as ’n aanvulling of ’n vervanging van die VVI, as voorspellers van persoonlike verbruik. Die vooruitskattingsvermoë van mediasentiment-indekse word beoordeel in ’n voorspellingswedren met die VVI. Die resultate van die voorspellingsoefening het aan die lig gebring dat die insluiting van media-gebaseerde sentiment-indekse as voorspellers in ’n model die voorspellingsfoute vir persoonlike verbruiksbesteding kan verminder. Die voorspellingsfoute het verminder in beide die korttermyn en langtermyn (tot en met twee jaar). Die uitslae bevestig die teorie dat sentiment in die nuusmedia inligting bevat oor die toevallige en vooruitskouende toestand van die ekonomie wat nie in die VVI vasgelê is nie. Die laaste bydrae beklemtoon die doeltreffendheid van die gebruik van domeinspesifieke woordeboeke wanneer analiste probeer om sentiment vas te lê. Domeinspesifieke woordeboek is outomaties opgestel deur gebruik te maak van masjienleertegnieke. Hierdie domeinspesifieke indekse vang die ekonomiese sentiment, wat deur verskillende algemeen gerapporteerde vertrouensindekse voorgestel word, meer akkuraat vas as wat die tradisioneel gebruikte Loughran-woordeboek kan. Hierdie raamwerk vergemaklik die proses van subjektiewe inligting-onttrekking uit media in 'n sentimenteboek, en handhaaf ook 'n vlak van deursigtigheid waartoe woorde vervat is in die konstruksie van die sentiment-indeks.

Gesamentlik bied die resultate wat in hierdie proefskrif aangebied word, aanvanklike ondersteuning vir die gebruik van teksanalise in Suid-Afrika as 'n alternatiewe manier om sagter ekonomiese aanwysers soos ekonomiese sentiment vas te lê.

Acknowledgements

This research was financially supported by the Bureau for Economic Research at Stellenbosch University (BER) and 71Point4. I would also like to thank the unnamed colleagues at the BER and Stellenbosch Department of Economics who provided insight and expertise that greatly contributed towards the research. Furthermore, without the team at News24 and Sabinet providing data, this research effort would not have been possible.

To fully acknowledge all parties that contributed towards this dissertation could be a thesis on its own. Inspite of this enormous task, I would like to personally recognize the role that family, friends and the Department of Economics at Stellenbosch have played.

Although this piece of work is the result of a project that started three years ago, it represents the culmination of a journey that I have walked with the Department of Economics at SU for over a decade. Unfortunately I am unable to mention each person individually, but I need to point out the monumental impact a single lecturer had on me - Nico Katzke. You changed the way I viewed economics and academics in general, and gave me a love and curiosity for the quantitative field of econometrics for which I will always be immensely grateful. You also bestowed on me a love for teaching that to this day drives me to inspire people the way that you inspired me.

I would also like to thank Prof. Ben Smit, Prof. Andrie Schoombee, Prof. Stan du Plessis, Gideon du Rand, Mike Lamont as well as the administrative staff at the department who no matter how busy they were, always entertained an open door policy.

I am grateful for the constant words of encouragement from my family, who in trying times were always supportive. To my mother - taking on this journey together with you has been amazing. Your determination to finish your PhD was an inspiration.

My colleagues at work, I cannot thank you enough for the laughs, the encouragement and understanding you have provided over this time. Amnon and Illana, no words are adequate.

To Johann and Monique - what an wonderous adventure this has been working with you. I know at times I could not have been the easiest student to work with, I thank both of you for your patience, guidance and wisdom. Without you this dream of mine would have never come to fruition. The lessons I have learned working with you extended beyond the realm of academics and I cannot put into words how much this has meant to me.

Finally, I would like to thank Melanie. Sharing a life with someone pursuing a PhD cannot be pleasant, but despite this you spurred me on through the tough times, gave me advice when I was lost, and love when I needed it the most.

Dedications

This thesis is dedicated to Johan, Rika, Maria and Melanie.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modelling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modelling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools. - Leo Breiman*

Breiman (2001b)'s perspective on the stagnation within the field of statistics has shifted over the past two decades. The statistical community no longer relies exclusively on data models, but today incorporates what is commonly known as machine learning alongside traditional statistical methods.[1] Recently, a similar shift has been seen to occur in economics. Although economics as a field was a late adopter of this shifting paradigm, large bodies of empirical work that embody

[1] With data models, the researcher starts with assuming a stochastic data model for the inside of the black box (ex. linear regression), while a data driven philosophy (or algorithmic culture) considers the inside of the box complex and unknown and as such the function $f(x)$ is unknown.

the data-driven philosophy is beginning to emerge.[2]

Economics as a field has greatly benefited from the use of automated machine learning algorithms. One such example has been information extraction from unstructured data sources such as text. Quintessential text analysis used to consist of the manual analysis of small text samples such as public announcements, selective company results, or monetary policy committee minutes. Analysing data in this manual manner is a labour-intensive process that can process far less material. Research that utilises computers for text analysis falls under a larger body of literature known as computational linguistics.[3] The incorporation of machine learning has increased the opportunities for research within the field of text analysis. This not only creates opportunity to utilise data from different sources such as news media, manager reports, or stock markets, but has been shown to provide information that can contribute to the improvement of forecasts of the economy, asset prices, and even the dynamics of inflation movements (Larsen and Thorsrud 2015; Chakraborty and Joseph 2017; Ardia, Bluteau, and Boudt 2019; Gentzkow, Kelly, and Taddy 2017).

1.1 Sentiment, text analysis and the economy

One of the first analyses to incorporate computational linguisticsto answer an economic question was Antweiler and Frank (2004)'s paper on the effect of market information and how it relates to volatility in the financial market. The paper empirically analysed 1.5 million messages from internet message boards from Yahoo!Finance and Raging Bull. The messages were all related to 45 companies that made up the Dow Jones Industrial Average at the time. The aim was to answer allegations from the financial press that internet stock message boards are able to influence the markets. Given that a large number of people who dedicate time in

[2] See Athey and Imbens (2019), Athey (2015), Athey (2017), Athey (2018), and Athey et al. (2019) for an extensive overview of how machine-learning techniques are being applied within economics as a tool for prediction and causal inference. Chapter 2 of this thesis also provides an in-depth literature overview of the topic and techniques.

[3] Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective. It also includes the study of appropriate computational approaches to linguistic questions such as sentiment.

their day to post content online (in a personal and professional capacity), this data should contain some information value. The findings confirmed a statistical relationship between the volume of posts in a given day and returns the following day. When the number of messages was above average, a negative return was observed the following day. The authors also found that including message volume as an explanatory variable improved volatility forecasts, as estimated by a GARCH(1,1) model. The paper was one of the first to incorporate such a large corpus of text data within financial economics. It illustrated that text could contain information on fundamentals in the market.

The use of text analysis to capture fundamentals that are hard to quantify can also be applied to macroeconomic indicators or proxies thereof. A growing body of literature is applying computational linguistics to construct consumer or business sentiment indices as a complement to the traditional survey-based methods (Daas and Puts 2014; Loughran and McDonald 2011; Van den Brakel et al. 2017). These indices offer a complement to survey-based methods as they do not suffer from decreasing response rates, cover a broad range of economic topics, and can relay information at a much higher frequency. Standard macroeconomic data is also usually released with a substantial delay, complicating the task of forecasting and making economic decisions within policy institutions and the private sector. Accurately predicting current economic conditions or key policy indicators of the economy, such as GDP growth, is hampered by revisions, asynchronous explanatory variables, and ragged-edge[4] issues.

To address some of these issues, nowcasting techniques have primarily focused on integrating high-frequency information predictors which are not subject to subsequent data revisions. Nowcasting models generally consist of both 'hard' and 'soft' information (Shapiro, Sudhof, and Wilson (2017)). Hard predictors are objective, quantifiable variables such as economic values of production, employment, and demand, while soft information captures subjective variables, typically collected through responses from surveys. These surveys usually aim to gauge economic sentiment from businesses and consumers.

[4] Due to different publication dates of economic variables, a problem such as ragged-edge arises. This results in an incomplete set of data for the most recent period, causing a ragged edge at the end of the sample (Wallis 1986)

Although the reason behind *why* consumer confidence and economic activity are highly correlated is still under debate, there is no refuting that attitudes contribute significantly to explaining future consumption growth (Ludvigson 2004). At the centre of the debate are two contradictory theories: the *informational view* and *animal spirits* (Barsky and Sims 2012). Information theory suggests that confidence indicators contain information that is not contained in other traditional macro variables and that is valuable for characterising future economic developments. Alternatively, the animal spirits theory, first introduced by Keynes (1937), suggests that independent changes in beliefs have causal effects on business cycles. However ambiguous the conclusion remains, the information view and animal spirits do not oppose the role of sentiment in business cycles or expenditure. The primary difference is that only the animal spirits theory implies causality.

In a seminal paper on consumer confidence and UK consumption expenditure, Acemoglu and Scott (1994) stipulate that unlike the error correction approach typically found in consumption expenditure, consumer confidence explicitly introduces a forward-looking element. This approach is in line with a standard theoretical approach to consumption: the Rational Expectations Permanent Income Hypothesis (REPIH) (Friedman 1957; Hall 1978). By including frictions within a rational expectations model, an observed increase in economic sentiment signals a potential higher future income. But given borrowing constraints the agent cannot consume today in anticipation of the increase in income. The agent delays consumption until the actual increase in income is realised. In such a theoretical framework, consumer confidence is said to predict higher future consumption. In the opposing animal spirits view, contemporaneous consumption decisions are a function of information about the future. Reacting to information from their environment, agents revise their economic outlook and increase discretionary spending. This revision is based not only on economic determinants such as prices and income but also their personal sentiment and expectations towards the economy. In the latter case, the psychological factors influence the perception of the economic environment and consequently the way in which the consumer responds to it (Shapiro 1972).

Irrespective of whether consumption is primarily driven by animal spirits or information theory, the influence of sentiment was highly evident in the periods following the Great Recession and sovereign debt crises in Europe in 2008 and 2011. Both industry and consumers delayed expenditure, adopting a watchful attitude towards an uncertain economy. It is in times of high uncertainty that a faster measure of conditional expectation of the economy, such as those shaped by news media, becomes valuable. Beaudry and Portier (2014) view this phenomenon as a way in which agents formalise expectation formation as a signal extraction problem. The authors argue that news driven business cycles are merely a result of economic agents adjusting their expectations based on information from various sources – the media playing a central role. Thus, economic boom and bust cyclical behaviour is a direct consequence of agents' incentive to speculate in order to profit. Examples of these boom and bust cycles can be found in the telecommunication boom (and the resulting bust due to underutilisation) at the start of the 1980s as well as the dot com bubble at the turn of the century. In spite of the simplicity of the mechanism through which the media-economic nexus operates, evaluating the true impact of media as an indirect driver in the economy remains a challenging task.

The first authors to reach a significant breakthrough in the dynamics of news media and the economy were Tetlock, Saar-Tsechansky, and Macskassy (2008). Their findings have acted as the foundation for most modern analysis concerned with text sentiment, both within economics and other fields. Firstly, their findings empirically support the hypothesis that the fraction of negatively associated words in media releases about an individual S&P500 firm helps to forecast lower firm earnings. The authors also show that there is an asymmetric effect where firms' stock prices briefly underreact to an increase in negative words. These findings jointly substantiate the notion that linguistic media content is able to capture characteristics of company fundamentals that are hard to quantify and not obvious at face value. This research encouraged other economists to consider how computational linguistics could be used in their domain-specific areas and, more specifically, which linguistic method should be used in the analysis. Today, text and sentiment analysis is one of the fastest-growing areas within this "big data revolution" and research.

This thesis applies text and sentiment analysis to answer three key questions relating to consumer confidence and the role of media based sentiment indices as complementary measures of consumer and business confidence. These indices are aimed at capturing current and future purchasing decisions and future business conditions respectively. Confidence indices also play a key role in capturing information about employment, GDP, and consumption expenditure not yet reflected in the official aggregate releases by governments. The confidence indicator is one of the key components in what is known as sentiment, the other is uncertainty. One of the key challenges that face operationalisation of these concepts is the fact that they are not directly observed (Santero and Westerlund 1996). Although measuring sentiment is not straightforward, survey-based indicators can be helpful in discovering agents' opinions on economic developments (Economic Co-operation and Development 2003). The advantage of survey-based measures lies in their ability to be representative of the population (consumer confidence) and use the opinions of key economic agents as inputs to the index (Girardi and Reuter 2016). In addition, survey data is also available earlier than most official statistics and is usually not subject to revision (ECB 2013).

1.2 Research questions and structure of thesis

In South Africa, the consumer and business confidence surveys are conducted every quarter to construct two indices: the consumer confidence index (CCI) and the business confidence index (BCI).[5] Although these surveys have a lot of advantages, they also have some weaknesses. The surveys come at a high cost and are dependent on the continued demand for the survey to be conducted.[6] Survey-based indices also face critique for other reasons. Surveys that consist of a few simple questions might struggle to capture a multidimensional economic concept (Commission 2008). Another criticism is directed at the construction of the questions. There is a risk that the meaning of certain concepts referred to in a question might change over time and, as such, interpretability of the index becomes difficult (Coertjens et al. 2012).

[5] See Binge (2018) for an extensive overview on the use of the BCI in forecasting macroeconomic variables.

[6] The questions in the confidence surveys are not collected in a dedicated questionnaire, but form part of a larger set of questions sent out to respondents

In light of these criticisms, this thesis addresses three research questions. The *first question* is whether a media-based sentiment index (MSI) constructed using text analysis and business news articles could feasibly replicate the current consumer-based confidence index. By applying sentiment analysis to hundreds of thousands of news articles spanning 17 years and constructing an index that encapsulates consumer confidence (through news media), I propose a South African media-based sentiment index as a possible complement to the consumer confidence index. Such an index can be constructed at a higher frequency, lower cost and in an automated fashion, while also capturing multidimensional aspects pertaining to economic sentiment that the traditional survey-based CCI fails to do. Survey questions are usually designed to capture confidence and uncertainty deliberately, through the choice of questions included. In the case of the media sentiment indices used in this thesis, however, there is no deliberate question design. In contrast, the dictionaries aim to capture a range of features of sentiment by using extensive word lists. The assumption is made that this high dimensionality will mean that the media index incorporates both confidence and uncertainty to some degree. While this could be tested separately too, that was not done in this thesis.

The *second question* pertains to the informational value inherent in media-based sentiment indices. I posit that the informational content of these media based indices compare favourably with that contained in the surveys. This hypothesis is tested in a forecasting horserace. Personal consumption expenditure and its subcomponents are all forecasted using a Bayesian VAR. Different modelling designs are compared to determine model which has the best out-of-sample forecasting ability at various horizons. The different model specifications include either no sentiment index, only the CCI, only the MSI, or both the CCI and MSI. The aim is to test whether media-based alternatives have any informational value above that which is contained in the CCI.

The *third and final question* relates to the technical construction of text-based sentiment indices in general. Until recently, almost all text-based economic sentiment indices have been constructed using what is known as the Loughran dictionary.[7] The core concepts of this dictionary originated in the accounting field. Given that literature has confirmed that domain-specific dictionaries do perform

[7] The definition and meaning of dictionary is discussed in detail in Chapter 2.

better than their general counterparts (Henry (2008); Labille, Gauch, and Alfarhood (2017); Loughran and McDonald (2011); Prollochs, Feuerriegel, and Neumann (2015)), is it possible to create an economic domain-specific dictionary that outperforms the Loughran dictionary? In addition, by employing machine learning tools such as Random Forests and Recursive Feature Elimination[8], is it possible to construct domain dictionaries in an automated fashion that exceed manually generated dictionaries' ability to capture economic sentiment across various confidence indices? All of these hypotheses are tested in the subsequent chapters.

The thesis begins with an overview of the data landscape and literature pertaining to the use of text analysis within economics and finance (Chapter 2). The chapter aims to introduce readers, who are unfamiliar on the topic of text analysis, to techniques often used, domain specific concepts and seminal papers. After the review of the literature, chapter 2 turns to a discussion of the data used within the remaining chapters of this thesis. Although the data is used slightly differently in each chapter, chapter 2 explores the data in a general manner, providing descriptive statistics, trends, discussions of text length and the sample period.

Chapter 3 begins with the construction of a proxy for news information using several news sources. The aim is for the index to be a complementary representation of the consumer confidence index constructed by the Bureau for Economic Research (BER) at Stellenbosch University through traditional survey-based methods. As previously mentioned, one of the benefits of news-based sentiment indices is the high frequency at which they can be constructed, while also incorporating a large array of topics that are discussed in the business press. In addition, the chapter also discusses a framework that allows researchers and practitioners the means with which to develop a monthly index in a semi-automated fashion. The method employs time series clustering and sentiment analysis to identify possible media sources (and linguistic methods) that contain information which closely resembles a traditional survey-based confidence index (CCI, in this case). While chapter 3 explores the degree to which it is possible to replicate the CCI using text-based indices, in chapter 4, the assumption that the CCI is the 'correct' measure of consumption expenditure, is relaxed.

[8] Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until an optimised number of features is reached.

In chapter 4, the aim is to empirically establish how well an MSI is able to forecast the outcome variable (PCE), compared to the survey-based CCI. To put it succinctly, how much information does the MSI contain about the future path of the PCE? The text-based sentiment indices are compared to the traditional CCI in forecasting future consumption expenditure (and its subcomponents) in South Africa. Using a Bayesian VAR framework, the out-of-sample predictive performance of the sentiment indices is compared for the periods 2007:Q1 to 2017:Q3 (43 out-of-sample quarters). The performance of each index is evaluated by applying density-based scoring models in conjunction with a statistical technique known as model confidence sets. Scoring models are used for forecast evaluation due to their ability to take into account the distributional properties of the forecast and not just the point realisation. This is especially useful when the forecaster wishes to incorporate uncertainty of the prediction as a measure of fit. Where many models are present and are producing vaguely similar outcomes, the problem of choosing a "best model" becomes challenging. Here, model confidence sets provide a statistical solution to narrowing down competing models up to a point where the difference in predictive power is negligible for the top-performing models. The results aim to indicate whether media sentiment indices have any informational value and, if they do, whether they are predictive or coincidental.

In chapter 5, machine learning is used as an alternative to the labour-intensive, manual process of constructing a domain-specific lexicon or dictionary. A framework is presented which operationalises the extraction of subjective information for economic time-series that aims to capture sentiment by using Random Forests. The application of sentiment indices within economics has increased recently, and it may be time to create a dictionary specially designed for application within economics, rather than the Loughran dictionary. Domain-specific dictionaries have been shown to better capture nuances within specific topics, such as politics or finance, where a lot of ambiguity might be present (Henry (2008); Labille, Gauch, and Alfarhood (2017); Loughran and McDonald (2011); Prollochs, Feuerriegel, and Neumann (2015)). However, the manual process used to create the Loughran dictionary is expensive, time-consuming, and potentially subjective. To address some of the weaknesses of the manual approach, an automated framework to identify key terms associated with sentiment in a specific domain is presented as a potential option to operationalise the dictionary-creation process. The model-generated dictionaries are empirically evaluated using out-of-sample RMSE as a measure of fit. A comparison is drawn between a well-known financial dictionary and the machine-generated indices. Another significant advantage the generated dictionaries have is that they include both unigrams and bigrams as features, uncommon in manually generated lexicons which usually only consist of unigram tokens.

These three chapters all pay tribute to the advantages and strides economics has made over the last few years due to the incorporation of data-driven modeling in the field of economics, and aim to contribute to this large body of knowledge by questioning the fundamentals such as dictionary generation.

Chapter 2

Text data and economics

Live with the data before you plunge into modelling. - Leo Breiman

Digital text footprints are capturing an ever-growing share of economic agent interaction, communication, and social culture. This vast information generation is playing an important complementary role to more classic economic data traditionally employed in research. It allows economists to capture how agents' actions, strategies, expectations, or sentiment endogenously change, by observing the patterns they create or the historical digital trails they leave (D'Orazio 2017). Text analysis mostly starts with mapping w words to W tokens that are a matrix representation of the text. This vectorisation could include other steps such as removing stop words, which involves dropping tokens that are commonly used but convey little information, such as 'the', 'and', or 'but'. Once the design matrix is constructed, the information contained in the text can be applied to answer predictive problems or calculate a measure of sentiment. The most commonly used vectorisation technique is known as the *bag-of-words* approach. In this method, the word order is ignored. Consider $\mathbf{w}_i$ to represent a vector with length equal to the number of unique words in a document and whose w_{ij} elements are the occurrence of word j in document i. For document i:

Dreams of war, dreams of liars, Dreams of dragon's fire

The bag-of-words representation after removing stop words and punctuation is $w_{ij} = 3$ for $j \in \{dreams\}$, $w_{ij} = 1$ for $j \in \{war, liars, dragon, fire\}$, and

$w_{ij} = 0$ for all the other words if a predefined vocabulary was supplied. This representation can be extended to include phrases, also known as n-grams. If we take the *bigram* of the illustrative text above, the representation is $w_{ij} = 1$ for $j \in \{\text{dragon fire}, \text{dreams dragon}, \text{dragos liars}, \text{dreams war}, \text{war dreams}\}$ as none of the bigrams are repeated. By extending the design matrix to include bigrams, a limited amount of dependence between words can be described. This is especially true in specific domains where tokens such as *economic recession* and *economic growth* inform us on the state of the economy. Modelling this relationship comes at a cost as the dimensions grow exponentially with the order of n phrases. The dictionary (bag-of-words) approach is especially useful in cases where the outcome v_i is not observed for any i. An example of this would be sentiment or uncertainty analysis. The latent outcome cannot be directly observed as would be necessary if one were to employ a supervised learning technique. A disadvantage of the dictionary method is that analysis is heavily dependent on the function that maps prior information from W to outcome v_i. This implies that a dictionary's informational content is heavily dependent on the tokens, W, that any dictionary consists of.

Within finance, news articles, SENS (Stock Exchange News Service), and social media have all been used to try to predict asset price movements and study the causal impact of new information on underlying assets. Text data is generated at a higher frequency, which means this information contributes towards adjusting intermediate expectations, while traditional indicators play an important role in shaping long-term expectations of the economy. The work of Tetlock (2007) and Antweiler and Frank (2004) can be regarded as seminal work on the use of text in a financial setting. These papers employed economic news and internet searches to quantitatively measure the interaction between the text and stock prices.

Other studies not only consider the potential for text to assist prediction, but also explore how the data can be used to infer casual relationships or act as priors for parameters in structural economic models. To quantify the cost of continuing racial animus in the United States, Stephens-Davidowitz (2014) used Google search data to show that areas with racially charged search rates were a robust negative predictor of Obama's vote share in the 2008 and 2012 elections. The research also found that the Google data showed a much greater loss in votes than the survey methods. Estimates were 1.5 to 3 times larger.

One of the features that distinguish text data from everyday economic structured datasets is that text is ‘inherently’ high-dimensional. To illustrate, if we take a sample of text documents, each consisting of w words, that has been drawn from p possible words, then the cardinality of the vector[1] of words would be represented by p^w. Lets take, 30 Facebook or Twitter messages. If each of these messages only consists of the 1000 most common words, the permutations for the different arrangement of tokens would have as many dimensions as there are atoms in the universe (Aggarwal and Zhai 2012). The consequence of working with such high-dimensional data means that traditional econometric techniques are not feasible and a new toolset that draws from computer science, machine learning, and computational biology has become more applicable within economic literature over the last 20 years.

By embracing these new methods and the alternative, many opportunities have opened for both macro- and micro-economic researchers to study.

2.1 Applications of text as data in economics

Although the use of computers has greatly expanded the use of text data in modern economics, applications of text analysis date back to as early as the 1930s.

In a paper read at a meeting of the Econometric Society and American Statistical Association on 31 December 1932, Cowles (1933) explored whether professional forecasters were any good at forecasting the stock market. His study focused on three different cohorts: financial services, fire insurance, and William Peter Hamilton.[2] Both financial services and fire insurance companies performed inadequately in forecasting the market and on average showed a return of 1.40 percent less than the average common stock over the period. Peter Hamilton’s performance was also notably underwhelming, achieving a result better than what would ordinarily be regarded as a normal investment return, but poorer than the result of a continuous outright investment in representative common stocks for this period (Cowles 1933). To make the predictions for Peter Hamilton, Alfred Cowles man-

[1] The number of elements in a set or other grouping, as a property of that grouping.

[2] William Peter Hamilton at the time was the editor for the Wall Street Journal and published, over a 26-year period from 1904 to 1924, forecasts based on Dow Theory.

ually categorised the published articles as being ‘bullish’, ‘bearish’, or ‘doubtful’. Modern-day principles show a large resemblance to the work of Cowles, using some kind of count of token C, to try to predict outcome V, albeit that in modern-day text analysis, techniques are evermore dependent on computers for performing the categorisation.

The use of text data to predict financial outcomes in the seminal work of Tetlock (2007) is one such example. By analysing media sentiment from the Wall Street Journal’s well-known “Abreast of the market” publication, he aimed to quantitatively measure the interaction between the media and stock returns. Tetlock (2007) converted the vectorised text into 77 different sentiment dimensions using the Harvard-IV psychosocial dictionary and then applied principal component analysis to reduce the dimensions into a single component that he calls the pessimism factor. This single, compressed media sentiment factor was used to forecast stock market activity over the period 1984 to 1999. The results showed high levels of pessimism in the media that predicted downward pressure on market prices followed by a reversion to fundamentals. The second finding of the paper was that unusually high or low pessimism predicted high market trading volume. Finally Tetlock (2007) argued that the effect of media sentiment is transitory, and that media sentiment acts as a proxy for investor sentiment or non-informative trading (unrelated to fundamentals). This is in line with the results finding that stock prices usually revert within a week and, as a consequence, the media acts as a sideshow containing no fundamental information on asset values (Tetlock 2007). On the back of Tetlock’s work, studies using dictionary-based methods have seen substantial growth. Gentzkow, Kelly, and Taddy (2017) presents an excellent overview on how text has been applied in various fields, while Chakraborty and Joseph (2017) gives a focussed perspective of how central banks are currently utlizing machine learning in the context of central banking and policy analysis.

Besides media, online platforms such as Twitter and Facebook generate large amounts of text data. Bollen, Mao, and Zeng (2011) analysed the text content of daily Twitter feeds arguing that the data generated could act as a signal for collective decision making. They used tools such as OpinionFinder and Google’s Profile of Mood States to investigate whether the public mood, as proxied by tweets, correlates (or predicts) the movement on the Dow Jones Industrial Average

(DIJA). Their findings report a significant correlation between the tweet sentiments and the DIJA as well as a reduction in the Mean Average Percentage Error (MAPE) by more than 6%.

Apart from dictionary-based methods, regression-based techniques using tokens as predictors are also being explored. To test whether regression techniques could possibly improve stock return forecasts, Jegadeesh and Wu (2013) estimate the response of company-level stock returns using each individual company's annual report. They showed that the appropriate choice of domain-specific words is at least as important as, and perhaps more important than, a complete and accurate compilation of the word list.[3] The authors propose a regression model where coefficients summarise the average association between the occurrence of a given word and the stock's subsequent return. The weighted approach achieved a better out-of-sample forecasting performance than dictionary-based indices from Loughran and McDonald (2011).

In more recent work, Manela and Moreira (2017) applied support vector machines to text in order to derive a news-implied market volatility measure from 1890 to 2009. By using penalised least square estimation, the authors identified a small subset of words that contain information that is useful for predicting turbulence in financial markets. Firstly, they highlighted the importance of government policy as well as the notion of war on the subsequent market stability. Secondly, the paper found that high levels of volatility, as implied by news media, forecasted high future stock market returns. The paper contributes to a better understanding of the transmission mechanism between news media and the financial market, and how an increasing mention of certain domain-specific terms could have a real effect.

A large part of the literature in text analysis deals with applications within the world of finance. There are other examples of where text has been applied to better understand central bank communication, improve nowcasting, as well as measure policy uncertainty. Communications released by central banks play a vital role in all economic agents' decision making due to their importance in setting public policies. For the last two decades, central banks around the world have become

[3] This limitation of non-domain-specific dictionaries is further explored in chapter 5 where this thesis develops its own framework to generate domain-specific dictionaries

more transparent in an effort to better achieve their policy objectives by detailing their own projections of the economy and targeted policy rates (Mishkin 2004). This increase in the transparency of central bank communication allows central banks to improve the predictability of their policy and, as a result, can affect long-term interest rates, usually beyond the control of the banks (Bernanke 2004; Woodford 2005). Lucca and Trebbi (2009) explored this idea by analysing the statements from the Federal Open Market Committee (FOMC). They show how applying textual analysis to the statements offers a quantitative interpretation of the information contained in the verbal or written content. In addition, the findings also reveal the important role central bank communication plays in affirming future policy decisions. Applying linguistic measures, the authors showed that short-term treasury bonds respond to unexpected policy rate decisions, while longer-dated instruments mainly react to the changes in the communications by the Fed. This is mainly due to the fact that long-term instruments are mostly driven by market expectations of future policy. This finding was also supported by a univariate model, where communication from the central bank was able to predict future policy rate actions with a lead of more than a year:

> By emphasizing the role of central bank communication, rather than the immediate setting of interest rates, our analysis highlights an important dimension of monetary policy that has received limited attention in the empirical monetary economics literature. - Lucca and Trebbi (2009)

Central bank communcations can also have an effect on financial stability. Born, Ehrmann, and Fratzscher (2014) investigate the effect of central bank communication sentiment and the repercussions it has on stock market returns and volatility. The paper empirically evaluates the reactions of stock markets to more than 1000 releases of Financial Stability Reports (FSRs) and speeches by 37 central banks over 14 years. Their findings suggest that FSRs play a vital role in stock market volatility and returns, while the verbal communications had modest impact over the sample. The interviews were only significant during the financial crisis, indicating the importance of verbal communication during times of stress as a policy tool to reduce volatility. The paper also draws special attention to the dynamics of central bank communication and the role that certain communication tools have to play in different times of financial stability.

2.2 South African news article dataset

This section presents the dataset of economic and financial news articles for South Africa discussed throughout the rest of this thesis.

Data was collected from three different sources: Meltwater, Sabinet, and News24. These sources represent a large readership throughout South Africa with the aim of capturing the general sentiment of the diverse readership. As a reference of sentiment we use the consumer confidence index (CCI) survey data from the BER. This survey will be compared with various constructed news-based sentiment indices. The analysis in chapter 3 is restricted to the period February 2009 to October 2017 as all datasets have overlapping information for this time frame, while chapters 4 and 5 use data from early 2000 up to 2017.

The data provided by Meltwater was extracted from an online platform using "Boolean Search" technology. Meltwater is a media-monitoring company that tracks various online media sources. In order to search through all the data, one has to construct search terms using keywords. All articles which adhere to the specified Boolean search criteria were exported to a CSV file at an aggregated level.[4] The exported files contained information on the original link to each article, the date analysed, the source, and most importantly, the sentiment as calculated by Meltwater's propriety classification algorithm. The searches were sub-divided into three categories: Consumer confidence, Business confidence, and Job-market. The keywords used to search the editorial content are aimed at extracting as much information as possible on the current economic environment. The confidence search terms aim to capture demand factors, while observing the sentiment in the job market could potentially give insight into the supply economy factors.

Figure 2.1 shows the different sources and the number of articles identified through the Boolean searches. In total, 207 314 articles were assigned a sentiment

[4] These searches are available in appendix A.1

score. Meltwater's dataset is unique in this analysis as the sentiment scores for the articles were already assigned without any intervention from the authors.

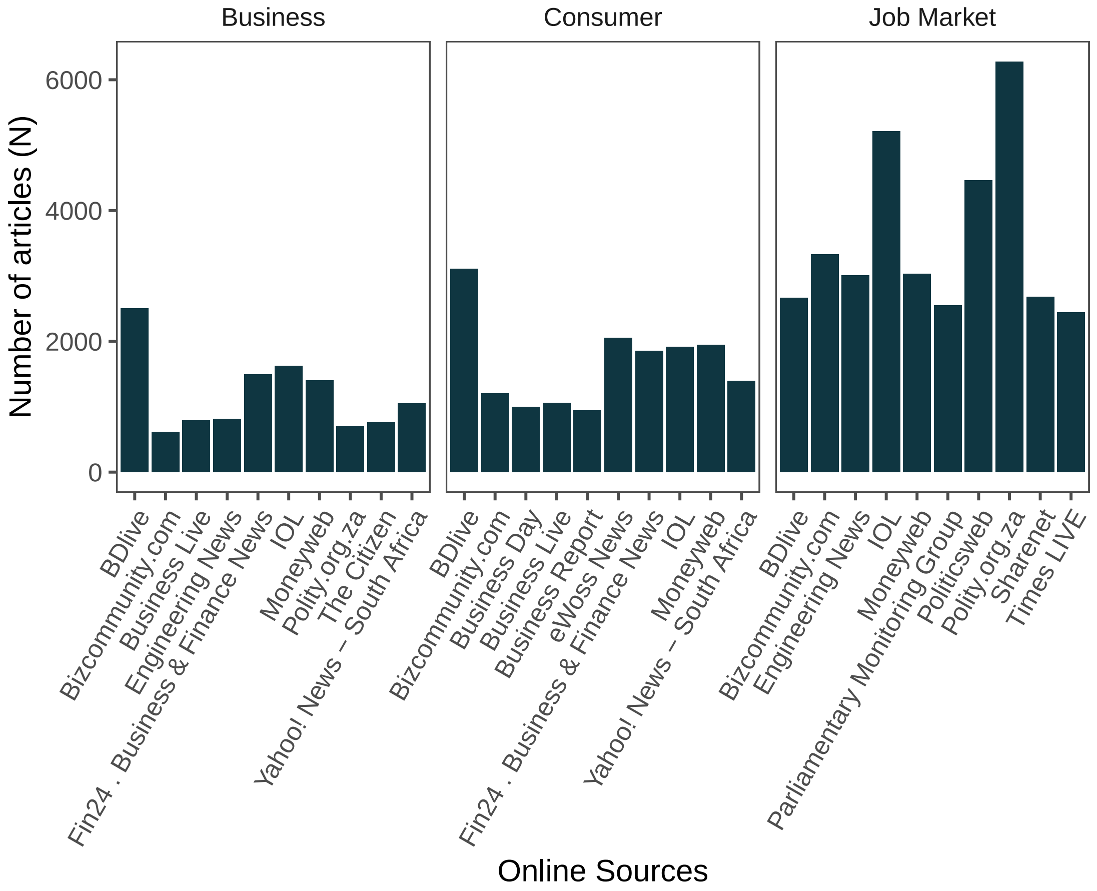


Figure 2.1: The online source of articles as identified thorugh Boolean searches using Meltwater's proprietary platform.

Our second source for news articles is News24. The data was received in raw text format. This allowed for a much deeper analysis of the text. The dataset consisted of around 1.9 million articles, spanning 15 years, from various media channels in the group. It also contained articles which were not in English. We made the decision to limit our analysis to English articles to avoid the complexity of having to translate non-English texts. The total number of English articles in the set was 1.2 million. To further reduce the computational burden, the sample was restricted to articles labelled as news, archived news, or financial news. This filtering of the data was done to restrict the data to relevant topics, removing some of the noise from sport or automotive publications. This, along with the restriction on the period of analysis, resulted in a complete dataset of 163 526 articles, of which 52 633 articles were labelled "financial". Figure 2.2 gives an indication of the number of articles per month that were being published online over the period of the investigation.

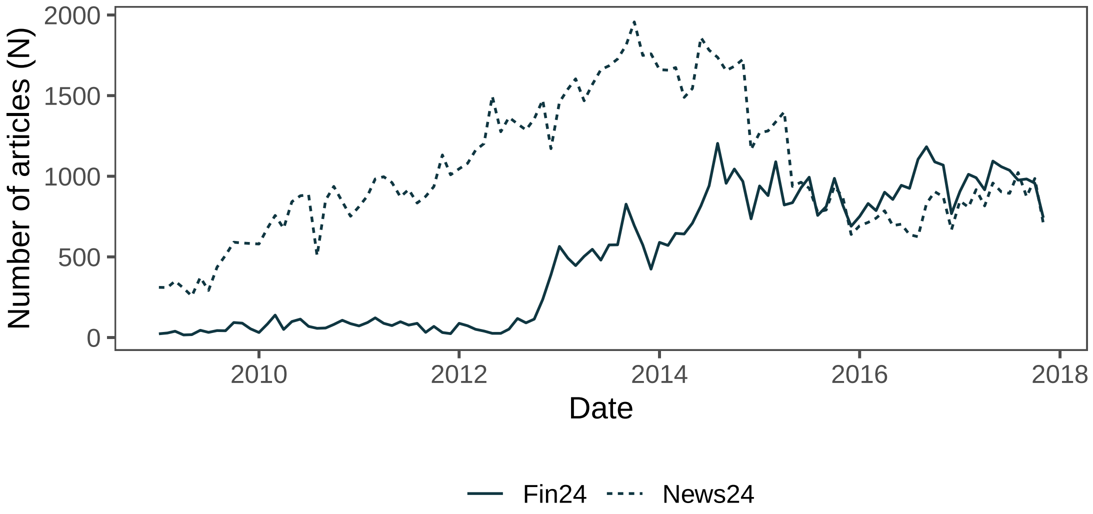


Figure 2.2: Monthly number of articles from January 2009 to October 2017 from News24 used in the construction of the indices.

It is clear that for articles with a "financial" tag, the bulk of the corpus exists after 2012. This is mainly due to the move by News24 towards online platforms and their promotion of a low subscription fee. The majority of the articles was classified "archive". Given that not much is known about these articles, it was decided that these should fall under the general "News24" descriptor, rather than being treated as financial news.

The Sabinet newspapers represent a large readership throughout South Africa. The aim of using these specific publishers is to capture the general sentiment through diverse readership. According to the All Media and Products Survey (AMPS) of 2015, the Sowetan is by far the biggest newspaper of the three in

Sabinet, having an exclusive readership of around 1 611 000. Business Day and the Financial Mail has a joint readership of 200 000. The Financial Mail especially has a key focus within the financial, investment, and political publications. All of these newspapers fall under a publishing house called the Tiso Blackstar Group which specialises in print and digital media products.

The PDFs received contained digital scans of article snippets from the respective newspapers. These PDF files were converted to text format using `pdftools`, a utilities library in R created by Ooms (2017). The package is based on the popular UNIX library called `libpoppler`, commonly used to render, extract, merge, and other utility features needed to augment PDF files. The completeness of the text extraction from the PDF files was highly dependent on the quality of the file provided. Given the nature of the problem, all news articles which contained less than 150 words after stop words were omitted, were discarded.[5]. Stop words are commonly used words such as *the*, *and*, *a*, and so forth that do not contribute anything towards an understanding of the content of the underlying text.[6] Another problem with this data source was missing observations for the period 1 January 2015 to 1 May 2015. In order to accommodate the construction of the indices, the data for the missing period was substituted with data from the “News24” dataset.

[5] This rule was also applied for the News24 data set.

[6] Stop words from the Bouchet-Valat (2019) library in R was used.

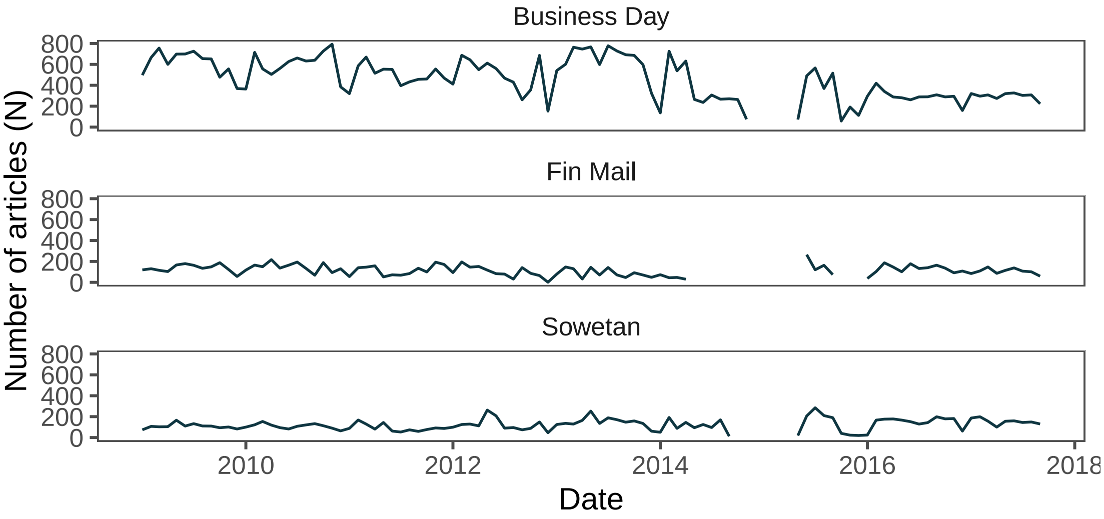


Figure 2.3: Number of articles from January 2009 to October 2017 as provided by Sabinet. A large proportion of the news articles is from Business Day.

The missing articles can be seen in the figure 2.3 as straight lines between the points for the period beginning of 2015[7].

Table (2.1) shows the summary statistics of the news articles used for analysis in each of the chapters.

Table 2.1: Summary statistics of articles used in each of the chapters

Chapter	Source	Articles (N)	Articles/month	Articles (SD)	First	Last
Chapter 3	Meltwater	207,314	643.83	499.99	2009-01-01	2017-12-01
Chapter 3	Business Day	46,427	459.67	195.88	2009-01-01	2017-10-01
Chapter 3	Fin Mail	10,294	115.66	49.15	2009-01-01	2017-09-01
Chapter 3	Sowetan	12,099	123.46	53.46	2009-01-01	2017-09-01
Chapter 3	Fin24	52,633	491.90	406.42	2009-01-01	2017-11-01
Chapter 3	News24	110,893	1036.38	422.17	2009-01-01	2017-11-01
Chapter 4 & 5	Business Day	122,819	33.06	15.10	2001-01-03	2017-10-02
Chapter 4 & 5	Fin Mail	24,529	35.19	16.14	2001-01-05	2017-09-28

[7] Upon a revised request to the data providers, the author was not able to obtain the missing data.

2.3 Summary

After starting with a brief explanation of how text is vectorised for most text analyses, the uses of text within finance and economics were discussed. The most commonly applied techniques make use of a dictionary-based method to either predict, classify sentiment, or quantify a measure uncertainty. More advanced techniques such as neural networks have also been successful in other domains of text analysis (Chen and Manning 2014; Goldberg 2016; Ginsberg et al. 2009). Using linguistic models with richer representations such as word embeddings has been less prevalent in economics and finance, but is showing promising results (Peng and Jiang 2015; Theil, Stajner, and Stuckenschmidt 2018; Jacobs, Lefever, and Hoste 2018). Following an extensive overview on text in economics, the chapter proceeds to introduce the novel data used throughout this thesis. The final data set consisted of media articles from three different sources: Business Day, Sabinet, and Meltwater, each showcasing a different narrative on the state economy. The amount of text was heavily filtered to decrease noise as well as ease the computational burden. Preference was given to articles pertaining to economic or finance topics, as is the custom with text analysis in this domain. In total, approximately 450 000 articles were used in chapter 3 to test the feasibility of each data set, while the subsequent chapters exclusively used the Sabinet data set due to favourite analytical results found in chapter 3 and the data being available from early 2000.

As the volume and velocity of online text information increase over the next couple of years, so will the importance of text within economic research. At the forefront of this growth is the rapidly expanding frontier of methods within the field of machine learning. This is supported by an equally fast-developing domain, cloud computing. This combination allows for fast estimation of methods that are calculation-intensive or where the volume of data is large – a problem often experienced in text analysis.[8]

[8] For an extensive overview, see the seminal work of Gentzkow, Kelly, and Taddy (2017) which provides an in-depth review on the use of text as data in the field of economics.

Chapter 3

Media based sentiment indices

> *Remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems. - Leo Breiman*

As the analogue era slowly fades into the "video-cassette" or "floppy disk" of yesteryear, the new digital age is generating information at an ever-increasing rate. Data is being generated at higher volumes and appearing in a host of different formats. Much of this information is a by-product of economic activity, rather than physical sampling or surveying. The private sector is developing innovative ways to generate and collect extensive datasets which can be converted into new revenue streams, without too much direct production cost – a "collect it if you can" mentality.

Economists are also embracing the data revolution, complementing these developments by rethinking how this kind of data could be employed within everyday economic analysis. New types of data are being generated faster and with far greater scope and coverage. One of the influential early examples of how online data collection on a large scale was used within economics to complement more traditional methods is the Billion Prices Project of Cavallo (2013). Under the BPP, prices from hundreds of online stores, spanning over fifty countries, are collected daily. In this way, it offers an alternative to traditional price indices with the advantage of being a real-time measure that is also available at a far higher frequency.

In this chapter, we construct and assess text-based indices that function as a proxy for economic sentiment found in online news media. In using news media data (as opposed to social media data) in the analysis, the chapter aims to overcome standard pitfalls such as personal privacy concerns, volume of text to analyse, as well as the difficulty of obtaining reliable historical data. We aim to contribute to the understanding of the feasibility of using these indices in an emerging market economy such as South Africa. The thesis is the first of its kind within an emerging market context. In emerging markets, especially South Africa with its diverse socio-economic landscape, the question of whether news can wholly act as an alternative consumer confidence indicator is an interesting one. For countries with a diverse socio-economic landscape, it could be argued that the opinion in the news is not a proxy for the attitude of the population as a whole. Despite this, the communications literature argues that news is an outcome of supply and demand. The journalist does influence the article (supply), but that is also a oversimplification. The journalists often view part of their role as representing the views of the public they report to (demand side pressures). In addition, the news articles are potentially influenced by the editor, specialists quoted in the article as well as other influential people (Reid and Du Plessis 2011). Consequentyly, by including a number of newspapers that represent a large readership of opinion leaders within society, it is quite reasonable to assume that the views in the newspaper are likely to come close to representing the views of society.

The survey-based business and consumer confidence indices developed by the Bureau of Economic Research (BER) of South Africa are some of the most widely quoted in South Africa. Their indices are based on the confidence index of the University of Michigan (UM). Research has confirmed that the confidence index developed by the UM has helped to forecast macroeconomic outcomes after having controlled for a host of economic factors such as disposable income and past personal consumption expenditure (Bram and Ludvigson (1997); Curtin (2007); Souleles (2004)).

A second contribution of the thesis is to suggest a framework by which researchers and practitioners can develop a monthly index that could act as a potential complement to the traditional survey-based consumer confidence index using multiple online media channels and dictionaries. Sentiment analysis forms part of a larger field called computational linguistics and although consumer sentiment consists of confidence and uncertainty, in this thesis sentiment refers to the constructed sentiment score or polarity as derived from computational analysis. These indices can be incorporated as complementary or alternative measures within national statistics as an indicator of consumer sentiment. By incorporating daily news and editorial content, the index aims to capture market information from a plethora of different channels, embodying a shared view of economic agents (as represented by the authors of the media articles) that extend beyond the opinions of select professionals.

This unstructured information set offers higher dimensionality and higher frequency than survey-based methods. The data allows for a bottom-up modelling approach where individual data points (news articles) are combined to model aggregate economic fluctuations. A time-series clustering technique is used to identify which of the subset of indices created (each relying on different combinations of dictionaries and online news sources) best reflects the current BER consumer confidence index (CCI). The aim is to identify the data sources and dictionaries that most accurately mimic the traditional survey-based index. This relationship is statistically tested by evaluating whether the MSI and the CCI is cointegrated, as well as testing for Granger causality.

The clustering framework[1] presented in this chapter provides a scalable solution to the high dimensionality problems associated with analysing text data from different sources and multiple methods of analysing its content. This chapter identifies candidate indices that could easily be analysed in a more robust statistical manner. The results from the cointegration and Granger-causality test indicate that there exists a contemporaneous and leading relationship between the newly constructed media-based sentiment indices and the survey-based CCI. These results pave the way for future research into how alternative datasets and especially text data can be incorporated to complement more established economic indicators.

The rest of this chapter is structured as follows: an overview of the current literature on consumer confidence and how to measure it is presented in section

[1] The framework pertaining to the construction and clustering of different sentiment dictionaries and data sources.

3.1. This is followed, in section 3.2, by an explanation of how computational linguistics and the bag-of-words approach can be used to extract sentiment from a piece of unstructured text. In section 3.3, we present the methodology adopted to construct the indices and a description of how smoothing techniques are employed to account for the sporadic nature of the constructed series. The section ends with an explanation of how we use dissimilarity measures and clustering to construct different indices and how these results feed into multiple media-based sentiment indices (MSI) that are constructed using principal component analysis. In the final section, we discuss the composite media sentiment indices and consider whether the media can be used in this way as an alternative index to measure consumer confidence.

3.1 Understanding consumer confidence

Although the mechanism through which consumer confidence affects the general economy is still a continuing debate, two primary mechanisms have been suggested (Shapiro, Sudhof, and Wilson (2017)). The first is an innate inability to capture reactivity of economic agents in times of uncertainty:

> Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of *animal spirits*, a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. - Keynes (1937)

The ‘animal spirits’ hypothesis first put forward by Keynes (1937) proposes that an unexpected change in the business cycle could occur due to the ‘gut’ (sentiment) of economic agents who respond to subjective foresight rather than quantitative evidence. This, in turn, changes economic activity through a consumer sentiment shock. These shocks are not uncommon – Blanchard (1993) explored the causes of the 1990 to 1991 recession by trying to isolate the causal sentiment variable.[2] Angeletos and La'O (2013) formalised the question through a rigorous mathematical representation of extrinsic movements in market expectations through what is labelled as sentiment. The formulation is one that requires neither a departure from rationality nor the introduction of multiple equilibria. The authors relaxed the assumption that economic agents should have expectations that match the actual state of the economy and in doing so, illustrated co-movement between economic activity and market expectations in the presence of a 'sentiment' or mysterious 'demand' shock. A mathematical formalisation of aggregate business conditions being affected through sentiment was derived by Benhabib, Wang, and Wen (2015). They employed a simple Keynesian framework showing that when consumption and production decisions are made separately by consumers and firms who are uncertain of each other's plans, the equilibrium outcome can indeed be influenced by animal spirits or sentiments, even though all agents are fully rational.

The second channel proposes that consumer confidence affects the business cycle through information contagion. The hypothesis is that informational news about the future state of the economy has already been internalised by economic agents, while not yet being captured in hard statistics. The reasoning behind the hypothesis is based on the early work of Pigou (1927) that states that its a shift in capital due to the difficulty of forecasting future demand that drive business cycles.

Beaudry and Portier (2014) and Barsky and Sims (2012) argue that only a limited number of unexpected business cycle fluctuations can be attributed to 'animal spirits', stating that uncaptured fundamental news is the primary channel through which sentiment affects subsequent economic activity. This is also true when investigating the medium-term effects of sentiment and confidence on output, technology, and investment. Barsky and Sims (2011) found that the information or 'news' channel has extensive explanatory power in the medium term, but evidence of playing a major part in the 'boom-bust' cycle, often cited in literature, is wanting. This state-dependent effect of consumer sentiment is further explored by Ahmed and Cassou (2016) who support informational contagion as

[2] Causal alluding to an economic variable which Blanchard (1993) states to be a proximate cause or decrease in relation to its normal determinants.

the main argument in times of economic expansion and 'animal spirits' in times of contraction.

For both schools of thought ('informational contagion' or 'animal spirits'), the consensus remains the same: consumer confidence and economic activity are strongly correlated and it is thus useful to incorporate consumer confidence into any model that wishes to forecast the future state of the economy. The reasons for the correlation are still debatable and remain open for discussion.

The concept of consumer confidence originated in the mid 1940s with George Katona at the University of Michigan (UM). The traditional approach to measuring consumer confidence is to construct an index by surveying economic agents using probability sampling for finite populations. The aim of the survey is to gain insight into the prevailing economic climate to have a quantitative method of incorporating consumer expectations into spending and savings models. The UM index is constructed by telephonically conducting 500 surveys on a monthly basis. The survey consists of fifty core questions and is constructed as a normalised sum of relative scores.[3]

In South Africa, a consumer confidence survey is conducted on a quarterly basis by the Bureau for Economic Research (BER) of South Africa.[4] The history of the index dates back to 1975 when the index solely consisted of the white population group, with black and other racial groups being included in the survey in 1982 and 1994, respectively (Kershoff (2000)). The survey result is the outcome of an area-stratified probability sample of 2500 households across South Africa. The survey is conducted on behalf of the BER by AC Nielsen/MRA with coverage in both urban and rural areas. For the majority population groups, white and black, sampling is conducted in metropolitan areas, cities, towns, and villages, while for the Coloured and Indian population, the surveyed area only includes the major metropolitan area. The stratified sampling aims to achieve a coverage of 92% of the urban adult population and 53% of the total adult population (Kershoff (2000)).

The interview is conducted in the home language of the respondent by a trained,

[3] This entails subtracting the percentage of negative responses from the favourable answers. The index was re-based to have an index value of 100 in December 1964.

[4] The BER is the only institution in South Africa that conducts consumer confidence surveys on a regular basis.

experienced fieldworker who is assigned a structured questionnaire that is directed at the head of the household. To ensure integrity of the survey, a minimum of 20% validation check is performed in order to validate the work of each interviewer.[5]

The questions used to assess consumer confidence are:

1. *How do you expect the general economic position in South Africa to develop during the next 12 months? Will it improve considerably, improve slightly, deteriorate slightly, deteriorate considerably or don't know?*
2. *How do you expect the financial position in your household to develop in the next 12 months? Will it improve considerably, improve slightly, deteriorate slightly, deteriorate considerably or don't know?*
3. *What is your opinion of the suitability of the present time for the purchase of domestic appliances such as furniture, washing machines, refrigerators etc. Do you think that for people in general it is the right time, neither a good nor a bad time or the wrong time?*

Although this chapter only focuses on the consumer confidence index, the BER also conducts research on business confidence (Business Confidence Index) in South Africa. South Africa was one of the first seven countries to start conducting qualitative assessment of the business environment using the German-based Ifo method back in March 1954.[6] At the end of each quarter, senior executives from the trade, manufacturing, and building sector complete a questionnaire with a small number of questions. With each round of surveying, the questionnaire is sent out to the same executives in each sector, thus ensuring a panel is established. The number of surveys sent out is 3800 in total, distributed as 1400 in the building sector, 1400 in trade, and 1000 in manufacturing. Both these indices have been proven to be very good leading indicators, be it for business cycles in the case of the BCI or consumer spending/savings for the CCI (Khumalo 2014; Kabundi, Nel, and Ruch 2016; Ndou, Gumata, and Ncube 2017; Venter 2019).

[5] The validation check is done either in person or telephonically.

[6] The Ifo method is currently applied in 57 countries, of which some of the best-known surveys are those of the European Union and the Tankan in Japan (Kershoff 2000). See Abberger (2006) for a review and evaluation of the method and its reporting.

3.2 Sentiment through textual analysis

Computational linguistics is best known for its sentiment and topic analysis toolset.[7] A body of text can typically be characterised by examining two facets within the text: (1) the degree to which the text exhibits emotion compared to a neutral stance, and (2) the degree to which a certain emotion is deemed to be dominant in the writing. The psychology literature usually divides these emotions along two dimensions: valence and arousal. Valence captures the intrinsic goodness (positivity) or averseness (negativity) towards a subject, object, or body of text; arousal describes and measures the intensity of the emotion.

Recent case studies on the construction of consumer confidence indices using online media data are starting to make their way into the economic literature. These 'sentiment indices' are constructed using text-based analysis. Two obvious advantages of text-based measures of economic tracking are the coverage and cost aspects thereof. Primary research such as surveys are inherently expensive to conduct, and can potentially be subject to small sample bias (Ludvigson (2004)). Contrarily, text-based indices are fast to adapt, cover a wide range of topics, and their construction can be automated programmatically. These text-based indices do, however, face some challenges.

Survey based methods are design to be representative of a specific population. As the underlying data source and journalists change, the validity and representation of a single newspaper may have a significant impact on the index itself. In addition, circulation and readership numbers of the news sources that make up the index. The indices also face a substantial initial cost in constructing a fully automated pipeline to analyze all of the text information and create the index.

In addition, there are choices that need to be made when constructing the media-based sentiment indices which are not insignificant. One of two (or a combination of) approaches are generally followed to quantify the sentiment of a body of text. The first approach, known as the "bag-of-words" approach, uses predefined dictionaries which consist of words associated with different emotions. An early reference to bag-of-words in a linguistic context can be found in Harris (1954)'s

[7] To read up on how topic analysis is being applied within the field of economics and finance, see Hansen and McMahon (2016), Larsen and Thorsrud (2015), or Hansen, McMahon, and Prat (2014).

paper that questions whether language has some distributional structure. Today this technique is taught as part of any computational linguistics course and widely applied accross multiple disciplines (Aggarwal and Zhai 2012).

Each of the dictionaries have their own bag of associated words and can thus deliver very different results when estimating a sentiment score. Loughran and McDonald (2011) recommends the use of a dictionary specifically designed for use in a financial or economics context. When constructing a sentiment score, the argument for the use of a financial dictionary, as opposed to the more commonly used Harvard Psychosociological Dictionary (Harvard IV), is due to negative connotations of words like tax, capital, expenditure, risk, etc. which are neutral within a financial context. Table 3.1 illustrates the bag-of-words concept by showcasing associated words and emotions from five different dictionaries. These dictionaries are used in this chapter to construct the sentiment indices:

- Loughran (Loughran and McDonald (2011))
- Harvard
- Henry's Financial dictionary (Henry (2008))
- Bing (Hu and Liu (2004))
- NRC (Mohammad and Turney (2013))

Table 3.1: Example of associated words and emotions from five different dictionaries. The associative positive and negative words can be used in a bag-of-words manner to derive various sentiment indicators.

Dictionary	Word	Association
Loughran	committed	constraining
Loughran	motions	litigious
Loughran	dangerous	negative
Loughran	assured	positive
Loughran	putative	superfluous
Loughran	predicting	uncertainty
Henry	under	negative
Henry	leader	positive
Harvard IV	expedient	negative
Harvard IV	repentance	positive
Bing	scandalize	negative
Bing	promoter	positive
NRC	revoke	anger
NRC	risk	anticipation
NRC	dispose	disgust
NRC	raptors	fear
NRC	soothing	joy
NRC	lowest	negative
NRC	pick	positive
NRC	specter	sadness
NRC	sunny	surprise
NRC	watchman	trust

We can see for instance that the word *assured* as in the Loughran dictionary has a positive connotation to it. This follows in the NRC dictionary where *lowest* has a negative connotation relating to it. Thus, it is important to note that the bag-of-words method does not take context into account, but is purely a one-to-one matching of word and association vectors.

The second approach that text analysis employs is machine learning algorithms, more commonly known as Natural Language Processing (NLP) methods. NLP is different from the basic association method as it takes into account the content and structure of a body of text. This entails the model having been trained on a large corpus of text in a supervised context where the predictor variables consist of a mapping between utterances and emotions. This approach towards sentiment classification is also known as a model-based approach. Using a model-based approach has the advantage of incorporating both the lexicon aspect of text analysis as well as introducing human intelligence into the model. Although machine learning algorithms can sometimes improve overall classification of sentiment, the algorithm is only as good as its training set. The models can also sometimes build sentiments with a confidence distribution, but due to the abstraction of the inner workings of the model, the inference dimension of the analysis can be difficult to explain. The models may also be retrained, which means that the sentiment scores could be different across different models.

One of the first papers to investigate the feasibility of constructing a social media sentiment index using the bag-of-words method is that of Daas and Puts (2014). Their approach considered the construction of a Dutch social media sentiment index (SMI) derived from Facebook, Twitter, and various other online data sources. The paper found that a strong association between consumer confidence (measured through surveys) and online sentiment (using computational linguistics) displayed by public Facebook messages. Their findings are consistent with the notion that a change in consumer confidence and a simultaneous change in online sentiment are driven by the same underlying phenomenon.

Van den Brakel et al. (2017) further investigated the concept of using social media to derive sentiment. They analysed a Dutch SMI using a multivariate structural time-series approach to estimate whether the inclusion of social media in the production of Dutch administrative statistics improved their accuracy. The paper also investigated the question of whether alternative data sources can be seen as complete substitutes for the traditional survey techniques. By estimating whether the two time series are cointegrated, the authors argue that a statistically significant outcome provides evidence that the different data sources are generated by the same underlying evolutionary process. Thus, their findings suggest that the SMI can be seen as a substitute for the more traditional survey approach. Following the modelling procedure as seen in Harvey and Chung (2000), the results of Van den Brakel et al. (2017) indicate that the Dutch CCI and SMI do indeed follow a similar evolutionary process, albeit that the underlying data generation differs.

Fraiberger (2016) turns to news information to infer economic sentiment. Fraiberger (2016) used a combination of Loughran and McDonald (2011)'s and Young and Soroka (2012)'s dictionaries to construct a sentiment index from the full corpus of economic news articles, produced by Reuters, across 12 countries over a 25-year period. The paper found that the constructed indices not only tracked GDP at country level, but contained information on future GDP growth which was not captured by consensus forecasts. A dictionary-based bag-of-words approach to sentiment mining is widely used, partially due to the ease of use and also due to the consistency of the dictionary over time. Dictionary methods can also be used to capture uncertainty and are not just used within the context of sentiment analysis. Baker, Bloom, and Davis (2016) developed an economic policy uncertainty index using a lexical-based method that identifies and counts articles containing the word "uncertain" and "not certain" and combines these with terms related to economic policy. Using human-verified readings of the articles confirm that the index proxies for movements in policy-related economic uncertainty.

In this chapter, both of these techniques are explored. In the case of the datasets where we have raw text, we calculate the sentiment scores using the five different dictionaries.[8]

3.3 Methodology

3.3.1 Creating a sentiment score

The total articles for a given dataset N_i^a where $i \in \{Meltwater_x, Sabinet_x, News_x\}$, a the individual articles and x represents the samples within each dataset. The total period over which the analysis was conducted is represented by $T^d = \{$2009-02-01:2017-09-31$\}$. As is standard with any analysis conducted on a corpus of this size, a data preparation step was introduced before analysis could begin. The Meltwater dataset already contained an associated sentiment score per article and thus did not need to be cleaned in comparison to the raw data that we received from Sabinet and News24. All data cleaning was done using R Core Team (2013)

[8] NLP scored sentiment data was acquired from Meltwater.

and the `tidytext` library by Julia Silge.[9]

The first step of cleaning raw text data is to remove all stop words. The lexicon we employed to remove the stop words contains 1149 stop words. A common second step of text analysis involves stemming. This leaves the core part of the word that is common to all of its inflections. For this thesis, we did not stem words, as the dictionaries that we used did not require this. The last bit of cleaning involves removing all editorial pieces with less than 150 words so as not to bias the dictionary method with low word-count articles. To construct a sentiment score, we identified the positive and negative words in each article $N_i^a \quad \forall \quad T$ using an external word list (dictionary) and doing a simple word count that consists of the positive plus negative words. We then normalised the count so that it reflected the relative proportion of positive and negative words within an article:

$$\begin{aligned} Pos_{i,t,n^a} &= \frac{PositiveWords}{PositiveWords + NegativeWords} \\ Neg_{i,t,n^a} &= \frac{NegativeWords}{PositiveWords + NegativeWords} \end{aligned} \tag{3.1}$$

where $i \in \{Sabinet_x, News_x\}$. The overall sentiment score for each article n_i^a, for $n_i^a = 1, \ldots, N_t^a$ at day t can then be defined as:

$$S_{i,t,n^a} = Pos_{i,t,n^a} - Neg_{i,t,n^a} \tag{3.2}$$

The polarity of the article is derived from the score. If the score of the article is greater than zero, then the overall sentiment for the article is deemed to be positive, and vice versa for a negative sentiment score. At this stage, the data set contained the Data Provider, ID, Date, sentiment score, and polarity of each of news reports. The index is constructed as the net balance of positive and negative articles within a month.

We created five different indices for each of our raw text datasets using each dictionary. Once the sub-sample indices were created, we also constructed source-level indices as an arithmetic average from the sub-samples. Figure 3.1 provides a visual representation of all the datasets used in the construction in the indices.

[9] Other packages we used as part of the data cleaning form part of what is known as the tidyverse (Wickham (2017)).

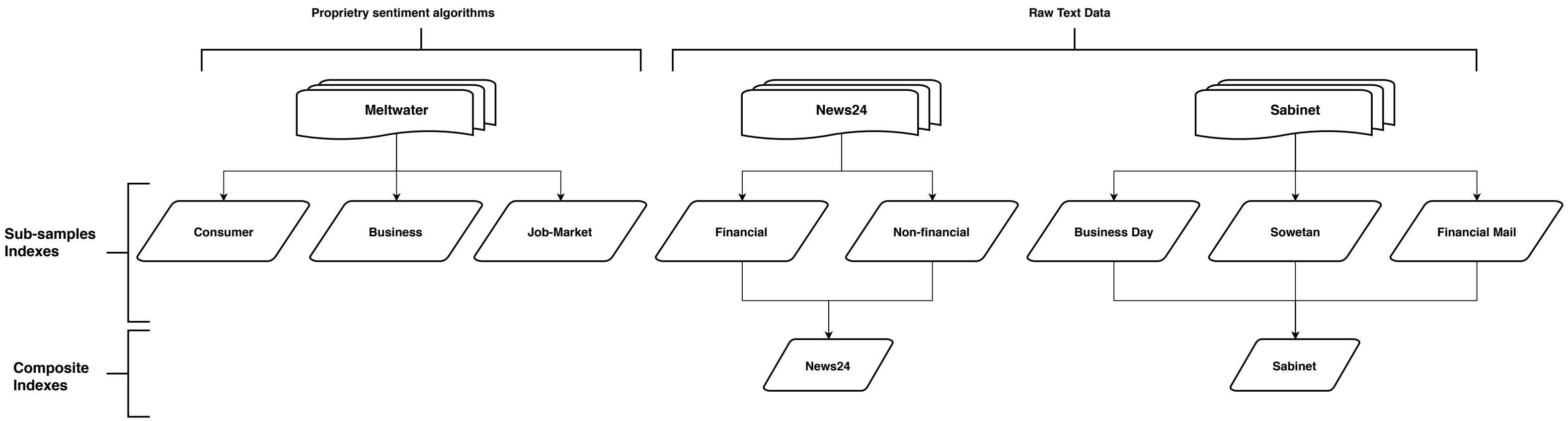


Figure 3.1: Data diagram of datasets used in analysis. The diagram illustrates how the different data sources were used or aggregated to construct the various sentiment indices.

3.3.2 Smoothing the sentiment index

Due to the volatile nature of the monthly indices, all series were smoothed using a Gaussian local level model.[10] The local level model is the simplest specification for signal extraction, only varying the level (u). If you were to believe that the underlying series had a cyclical component, then using a local level trend would be more appropriate.[11]

Let $\mathbf{y_t}$ denote a $N \times 1$ time-series vector of observations. The observations develop over time in terms of an unobserved vector ξ_t with $m \times 1$ dimensions, each at date t, for $t = 1, \ldots, T$:

$$\xi_{t+1} = \mathbf{T}_t \xi_t + c_t + \mathbf{R}_t \eta_t \quad \eta \sim N(0, Q_t) \tag{3.3}$$

$$y_t = \mathbf{Z}_t \xi_t + d_t + \epsilon_t \quad \epsilon_t \sim N(0, H_t) \tag{3.4}$$

Equations (3.3) and (3.4) represent the general linear Gaussian state space model that describes the dynamics of the system and are also known as the state and observation equations, respectively. The deterministic parameter matrices, $\mathbf{T}_t, \mathbf{R}_t$, and $\mathbf{Z}_t$, are of dimension $m \times m, m \times r$ and $N \times m$, with $\mathbf{R}_t$ being an identity matrix. Through the appropriate definitions of $\mathbf{Z}_t$ and ξ_t, certain unobserved structural components such as trend, seasonal, and cycle may be modelled. Vectors $\mathbf{c}_t$ and $\mathbf{d}_t$ are used to incorporate known effects about the expected value of the observations and states as including a dummy variable or explanatory variable with a fixed coefficient. For the purpose of this thesis, the latter drift vectors, $\mathbf{c}_t$ and $\mathbf{d}_t$, will be set to zero, to only capture the mean component. The model is estimated through the Kalman Filter's filter and smoothing processes.

$$E(\eta_t \eta_\tau') = \begin{cases} \mathbf{Q}_t & \text{for } t = \tau, \\ \mathbf{0} & otherwise \end{cases} \tag{3.5}$$

$$E(\epsilon_t \epsilon_\tau') = \begin{cases} \mathbf{H}_t & \text{for } t = \tau, \\ \mathbf{0} & otherwise \end{cases} \tag{3.6}$$

[10] This model is also known as a random walk plus noise state space model.

[11] An example of using multiple specifications of the underlying state space model van be seen in Van den Brakel et al. (2017).

Disturbances from the state and observation equations are assumed to be uncorrelated at all lags:

$$E(\eta_\tau \epsilon_t') = 0 \quad \text{for all } \tau, t, \ldots, T \tag{3.7}$$

as well as being independent from the initial state vector ξ_1. The initial state vector is assumed to be normally distributed with $m \times 1$ mean $\mathbf{a}_1$ and $m \times m$ covariance matrix $\mathbf{P}_1$:

$$\xi_1 \sim N(a_1, P_1) \tag{3.8}$$

The local level model is analogous to an exponentially weighted moving average (EWMA), with the added benefit of the variance and transition parameters (ϵ_t, η_t) being estimated through maximum likelihood. For the purpose of this study, diffused priors[12] will be used throughout. We believe a deeper discussion and exploration into the wide-ranging choice of priors are not warranted given the scope of this chapter and recommend reading Koopman and Durbin (2003) for a full mathematical discussion.

The model is estimated through the Kalman Filter's filter and smoothing processes. The purpose of the filtering mechanism of the Kalman filter is to update our knowledge about the state vector when new information about $\mathbf{y}_t$ becomes available to the system. Using the known distributional properties of the state, $\mathbf{a}_1$ and $\mathbf{P}_1$, the Kalman filter can be employed for the objective of estimating the conditional distribution of ξ_{t+1} for $t = 1, \ldots, T$ based on vector $\mathbf{Y}_t$ for the given information set $\mathbf{Y}_t = \{y_1, \ldots, y_t\}$.

The conditional distribution of ξ_{t+1} can be characterised by its mean, $\mathbf{a}_{t+1}$, and its covariance, $\mathbf{P}_{t+1}$:

$$\mathbf{a}_{t+1} = E(\xi_{t+1}|\mathbf{Y}_t) \tag{3.9}$$

$$\mathbf{P}_{t+1} = Var(\xi_{t+1}|\mathbf{Y}_t) \tag{3.10}$$

[12] This implies $a_1 = 0$ and $P_1 \rightarrow \infty$, as proposed in Koopman and Durbin (2003).

The mean of the conditional distribution, ξ_{t+1}, is obtained through an optimal estimator of the mean squared error matrix, $E((\xi_{t+1} - a_{t+1})(\xi_{t+1} - a_{t+1})'|Y_t)$, at time $t + 1$, $\forall\ \xi_{t+1}$.

Assuming then $\xi_t \sim N(a_t, P_t|Y_{t-1})$, it can be shown that $\mathbf{a}_{t+1}$ and $\mathbf{P}_{t+1}$ can be calculated recursively from $\mathbf{a}_t$ and $\mathbf{P}_t$.[13] After this forward pass whereby the recursive Kalman filtering process is applied to $\mathbf{Y}_t$, all information sets are stored. The state and disturbance smoothing recursive algorithm is then applied by proceeding backwards through all observations of the Kalman Filter output information set. State smoothing essentially estimates the state vector ξ_t based on the observation of the Kalman Filter:

$$\hat{\xi}_t = E(\xi_t|y) \tag{3.11}$$

$$V_t = Var(\xi_t|y) \tag{3.12}$$

with $\hat{\xi}_t$ as the estimated smoothed state and V_t as the smoothed state variance. Both the mean and variance of the state vector ξ_t are again obtained through backwards recursion. This leads to a smooth estimate, as will be seen when comparing the outputs to earlier defined techniques.

Figures 3.2 and 3.3 visually show the resulting smooth series derived from different data providers and dictionaries. The figure for the Meltwater data is seperated from the raw text sources due to the nature of construction.

[13] For a full mathematical derivation of this, see Mergner and Bulla (2008).

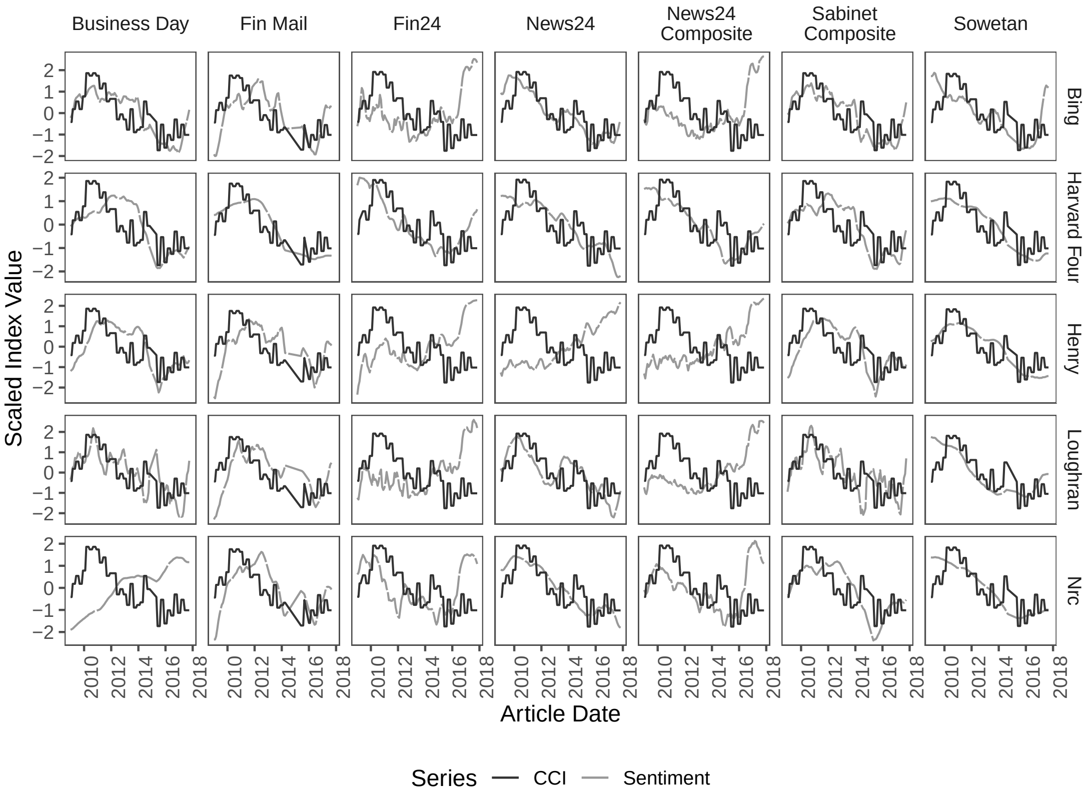


Figure 3.2: Comparison of media sentiment index and Consumer Confidence Index.

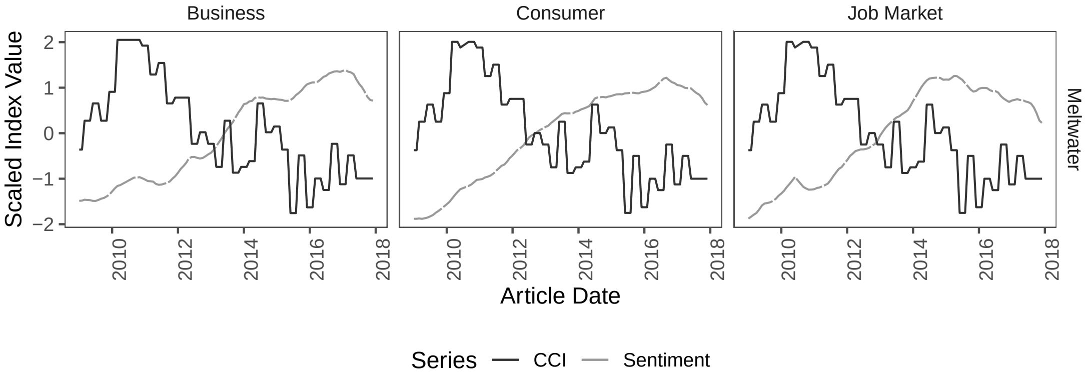


Figure 3.3: Comparison of media sentiment index and Consumer Confidence Index (Meltwater).

3.3.3 Quantitative analysis

Once all the indices had been constructed, we used time-series clustering techniques to better understand co-movement among the reference series, the traditional survey-based CCI, and the constructed media-based sentiment indices (MSIs). The aim was to sub-divide the large sample set through clustering into smaller homogeneous buckets to identify which of the indices created most resembled the BER's CCI by showing the lowest dissimilarity measures. Once the clusters have been identified, a composite index will be created from the cluster that contains the CCI. This newly created composite index will be compared to the CCI as a viable alternative.

Time-series clustering is an active research area with applications of the technique being seen in literature encompassing a wide range of fields. Although the technique is gaining traction within other fields, it is still underutilised within the field of economics. We employ this technique in order to help us to identify time series that behave in the same way. The hypothesis is that series that move similarly to the CCI would cluster in a homogeneous group. The steps for applying the analysis are two-fold. The first step is to construct an appropriate dissimilarity matrix between the MSI indices created and the CCI. Secondly, we used hierarchical clustering to analyse a large set of constructed indices to identify similar underlying patterns.

A key input in cluster analysis is determining a proper dissimilarity measure between two data series, where the main categories of the approaches can be divided into four groups: model-free measures, model-based measures, complexity-based measures, and lastly, prediction-based measures (Montero and Vilar (2014)). Each of these methods has its own strengths and weaknesses and, as such, a "best" dissimilarity measure will differ with each dataset. To choose the most appropriate distance measure, the first important step is to decide whether the clustering should be governed by shape-based or structure-based concepts (Lin and Li (2009); Corduas (2010)). When considering a shape-based approach, the main goal is focused on comparing geometric profiles of a series, while structure-based dissimilarity constructs aim to compare underlying data-generating processes (or structures).

For our study, it is important that the direction of change be the same, and because of this, we decided to use a non-model-based dissimilarity measure, called dynamic time warping (DTW). This technique was first popularised for time-series analysis by Berndt and Clifford (1994). The big advantage of DTW is that the frequency for the series does not need to be the same. The problem the analysis faces when choosing an appropriate distance metric is the time domain of the series that has to be clustered. All of the BER's indices are released on a quarterly basis, while the aim of the research is to create a monthly sentiment indicator.[14] This thesis also settled on this dissimilarity measure, as distance measures such as Pearson do a one-to-one correlation between X and Y. This would mean that one could lose out on possible leading time series when using methods such as Pearson correlation.

We start off defining r as a proximity measure and M representing all possible sequence of m pairs. The ordering of the observations is preserved by imposing monotonicity in order to avoiding futile looping as the algorithm assigns the closest points:

$$r = ((X_{a_1}, Y_{b_1}), \dots, (X_{a_m}, Y_{b_m})) \tag{3.13}$$

where $a_i, b_j \in \{1, \dots, T\}$ such that $a_1 = b_1 = 1, a_m = b_m = T$, where the distance between the coupled observations (X_{a_i}, Y_{b_j}) is minimised.

$$d_{DTW}(\mathbf{X_T}, \mathbf{Y_T}) = \min_{r \in M} \left(\max_{i=1,\dots,m} |X_{a_i} - Y_{b_j}| \right) \tag{3.14}$$

Dynamic time warping allows for the recognition of similar shapes between time series, even in the presence of signal transformation such as shifting or scaling. A toy illustration of how dynamic time warping creates a mapping between time series can be seen in figure 3.4 (Giorgino (2009)):

[14] For an extensive list on more complex dissimilarity measures, see Montero and Vilar (2014).

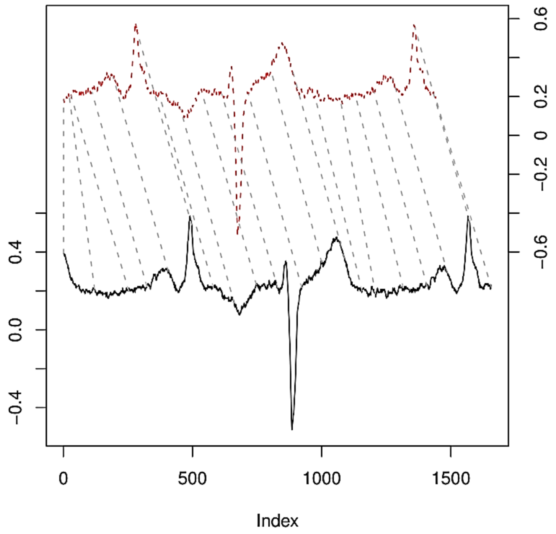


Figure 3.4: Illustration of how dynamic time warping is a mapping function between two points from (X_{ai}, X_{bi}) and not a one-to-one mapping at time t

Table 3.2 shows the result of the top 10 closest constructed online sentiment indices as per the dynamic time warping measure for both the CCI and the BCI. The results from the analysis confirm that the constructed indices are a better representation of the CCI than the BCI as per the lower DTW distance measure:

Table 3.2: Distance between CCI (BCI) and the constructed media sentiment indices as per the dynamic time warping distance calculation.

Reference Index	Sentiment Index	Distance (DTW)
CCI	News24 Loughran	44.91
CCI	News24 Bing	49.06
CCI	Business Day Harvard Four	49.13
CCI	Business Day Bing	50.49
CCI	Business Day Loughran	51.55
CCI	Sabinet Composite Harvard Four	51.59
CCI	Sabinet Composite Loughran	52.16
CCI	Business Day Henry	53.95
CCI	Sabinet Composite Henry	54.12
CCI	Fin Mail Harvard Four	56.06
BCI	Fin Mail Henry	59.05
BCI	Fin Mail Bing	59.96
BCI	Sabinet Composite Henry	61.33
BCI	Fin Mail Nrc	63.18
BCI	Fin Mail Loughran	64.97
BCI	Sabinet Composite Loughran	65.47
BCI	Business Day Henry	68.22
BCI	Business Day Loughran	70.54
BCI	Business Day Bing	72.19
BCI	Sabinet Composite Bing	75.80

Dynamic time warping is considered to be a measure and not a metric. This means that absolute numbers are not used to compare series and only relative comparison can be made. The results indicate that the Bing and the financial Loughran dictionaries tend to produce closer fits to the CCI. We also see that the sentiment indices constructed using News24 and Business Day data produce time series which best reflect the CCI based on the DTW measure of closeness.

From the total distance matrix we can conduct clustering. Within hierarchical clustering, there is a choice of two paradigms: agglomerative and divise. We use a commonly known hierarchical clustering method from R Core Team (2013), `hclust`, that involves creating clusters that have a predetermined ordering from top to bottom, also known as agglomerative hierarchical clustering. The algorithm starts by assigning each observation to its own cluster and then computes the similarity between each of the clusters, joining those that are most similar. This procedure is repeated until a final cluster is formed in a tree-like fashion. Both paradigms of hierarchical clustering possess what is known as a monotonicity property. This suggests that dissimilarity among clusters increase the higher up in the tree they merge. The height of the tree at each node is proportional to the value of the inter-group dissimilarity between its daughter nodes, while the individual observations are all plotted at height zero (Friedman, Hastie, and Tibshirani (2001)). This structure is more commonly known as a *dendogram.*

Before the clustering can start, a linkage criterion is needed to act as a function of the pairwise distances of observations in the dissimilarity matrix provided. We used the method of Ward Jr (1963) which calculates a merging cost when forming clusters A and B. Let $A = \{a_1, \dots, a_{n_a}\}$, and $B = \{b_1, \dots, b_{n_b}\}$ consist of observations in $\mathbb{R}^d$. Define the between-within, or e-distance $e(A, B)$, between A and B as:

$$
\begin{aligned}
e(A,B) =& \frac{n_A n_B}{n_A + n_B}\left(\frac{2}{n_A 1 n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(a_i, b_j)\right. && (3.15)\\
& \left. - \frac{1}{n_A^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_A} d(a_i, a_j) - \frac{1}{n_B^2}\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} d(b_i, b_j)\right). && (3.16)
\end{aligned}
$$

where n represents the number of observations in a given cluster and d the distance matrix between centroids of clusters. To state it more simply, Ward's minimum variance method aims to calculate the distance between cluster members and their respective centroid. The centroid of a cluster is defined as the point at which the sum of squared Euclidean distances between the point itself and each other point in the cluster is minimised.

3.3.4 Clustering results

Figure 3.5 shows the result from Ward's minimum distance hierarchical clustering implemented on a dissimilarity matrix calculated using dynamic time warping. The graph visualises the final choice of cluster. The tree was cut so that six clusters emerged. The decision of the cut was made based on the C-index, a cut criteria, of Hubert and Levin (1976) and is based on the ideology of "cophenetic" correlation.[15] This resulted in the CCI, Business Day (Loughran), and Sabinet Composite (Loughran) ending up in a single cluster.

[15] Cophenetic correlation determines how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points (Sokal and Rohlf 1962).

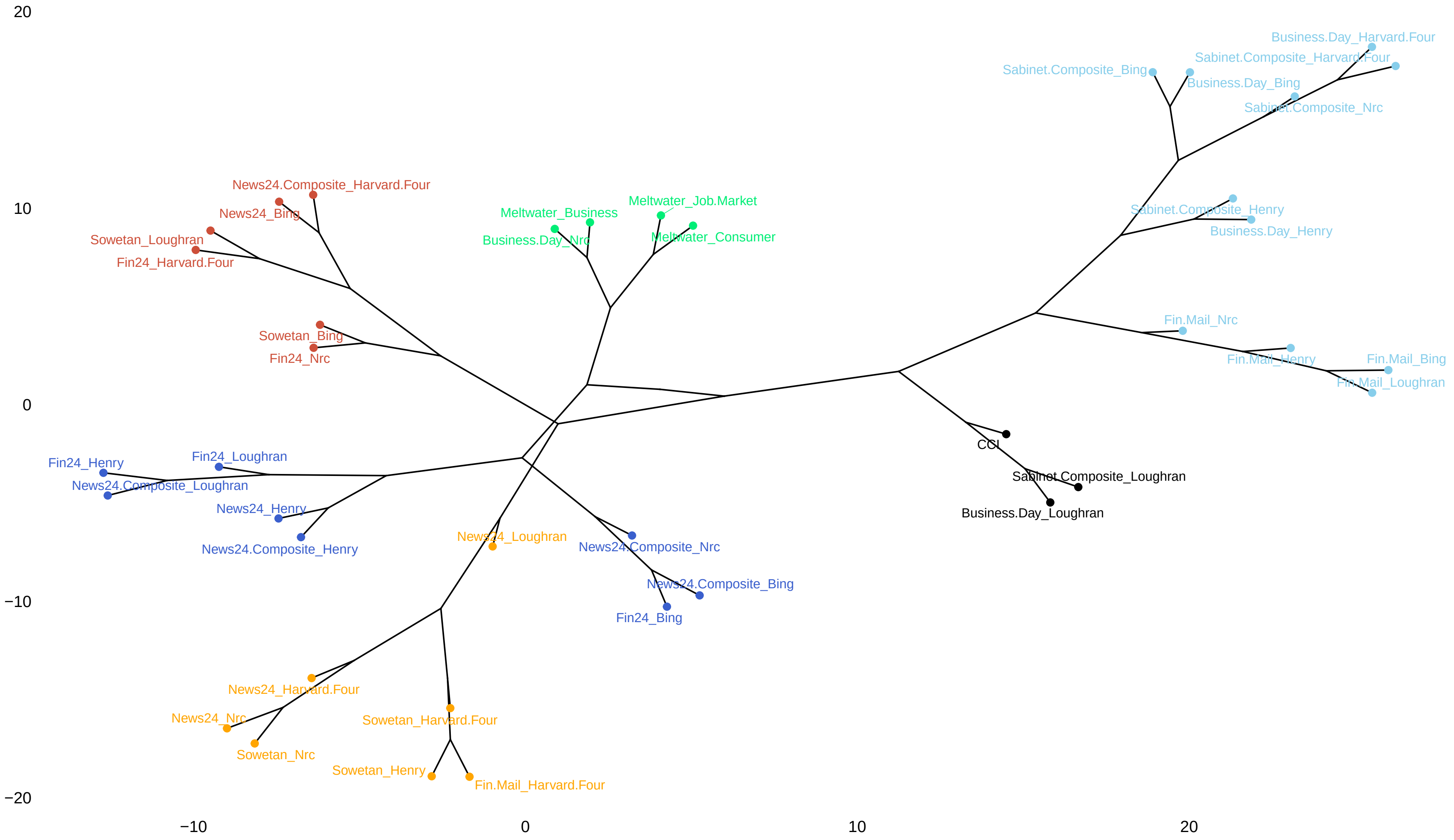


Figure 3.5: Visualisation of the hierarchical agglomerative clustering of dissimilarity measures between the media-based sentiment indices as measured by dynamic time warping.

The group that contains the CCI is dominated by the Sabinet news data. In terms of the dictionaries used in the analysis, Loughran and McDonald (2011)'s financially orientated lexicons delivered the best results.[16] Focusing on the branch that contains the CCI, the data source that produced an index most similar to the CCI was composite Sabinet and Business Day news data. Closely related to the cluster containing the CCI is the Financial Mail cluster. The Financial Mail's main readership consists of a select cohort in the population that has a key interest in the state of the economy, investment, and current political agenda. It is this specific editorial focus that most likely put the Financial Mail's cluster in close proximity to the CCI's cluster group. All of these constructed indices reiterate the important role that the financial dictionaries are playing in constructing a series that mimic consumer confidence on a monthly basis. Another characteristic to notice between the calculated series in the CCI cluster is the nature of volatility between points. News24, with a very large corpus, resulted in quite a smooth series, while Sabinet – of which Business Day is a subset – exhibited much larger changes between points in the series.

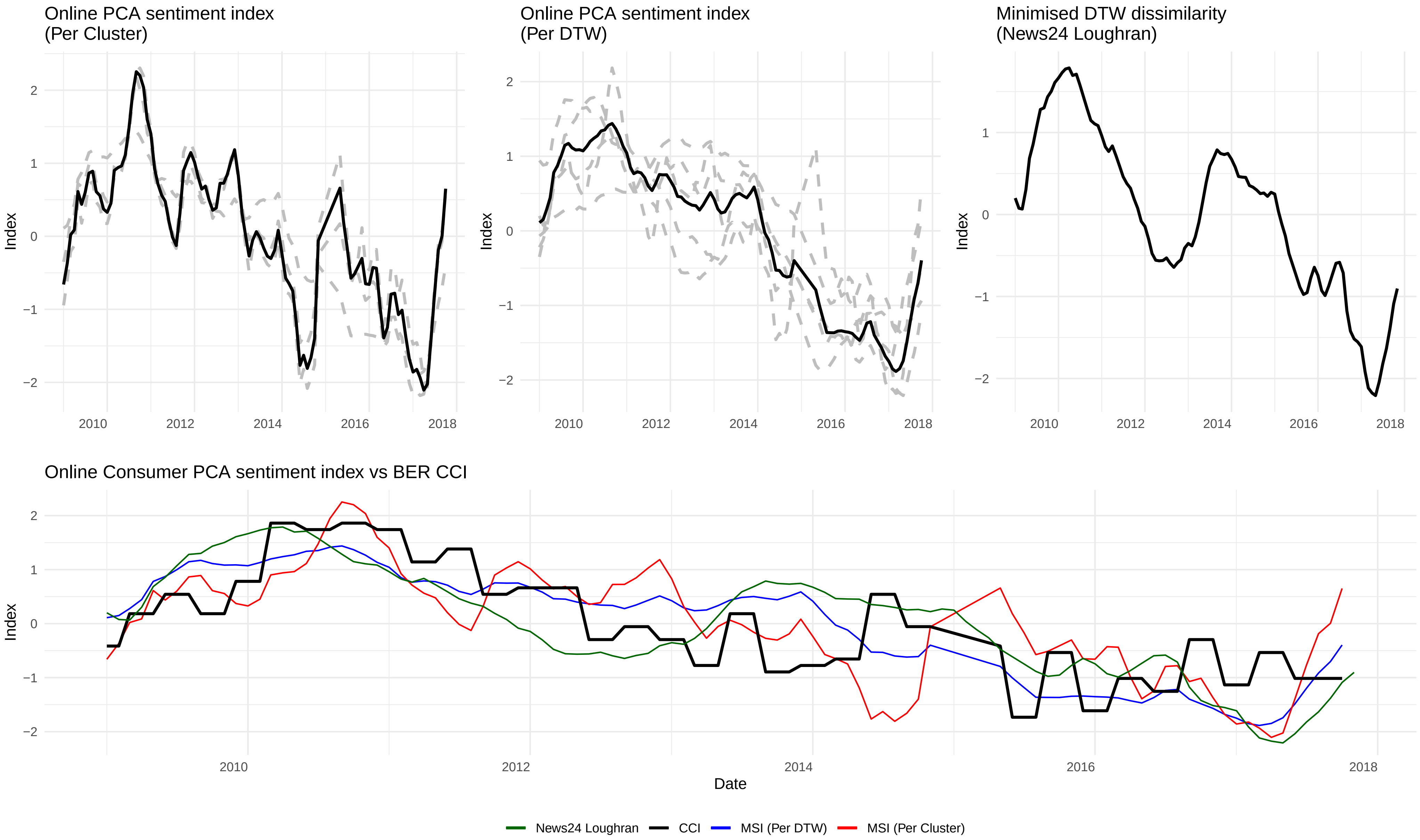
3.4 Discussion

The results from the clustering analysis reveal that there are clusters of indices that mimic the current survey CCI, even though the media sentiment indices all have a monthly frequency, while the survey-based indices have a quarterly frequency. This is an interesting finding on its own, given that the media-based indices do not ensure demographic representativity through careful sampling as is possible with some survey-based methods. Despite this disadvantage, the clustering approach does well in grouping indices that resemble the CCI. Three separate media sentiment indices (MSIs) were constructed for consideration: two composite indices using the clustering results, and one using only the raw DTW measure.

To construct the composite media-based consumer sentiment indices, principal component analysis was used. Principal component analysis allows one to use a set

[16] Although a clustering method cannot quantitatively be assessed, 'best' in the case of this analysis refers to a constructed series that best reflects the behaviour and shape of the reference indices.

of possibly correlated observations, $\mathcal{X} \in \{x_1, \ldots, x_N\}$, in an orthogonal transformation to convert them into a set of values of linearly uncorrelated variables called principal components. The first component of the analysis is considered to be the constructed sentiment index. Figure 3.6 shows the result of the constructed indices for the MSIs. The figure gives an indication of how the different constructions of a media-based index compare to the BER's CCI.



(Top left) Results of the PCA using series found in the CCI cluster. The red line represents the scaled first dimension of the constructed online sentiment indices. (Top middle) Results of the PCA based on the dissimilarity measure. The blue line represents index, while the dotted lines represent the underlying series used in PCA.. (Top right) The single index that minimised the distance between the CCI and itself as per DTW. (Bottom) Results of the three media sentiment indices overlaid on top of the BER CCI (black).

Figure 3.6: Composite indices constructed using PCA.

Table 3.3 presents the PCA results of the composite online consumer sentiment indices. The eigenvalues and percentage of Variance from PCA analysis is used to construct two different MSI indices based on the clustering results as well as just using the top five indices most similar to CCI (as per DTW). The results show that the first component captures the majority of the variation present in the time series. The cumulative percentage of variance is equal to 96.29% and 78.40% respectively. The composite index is constructed using the clustering approach was equally weighted. The index that incorporates the top five closest indices as per the DTW measure has almost equal weighting in the composite index.

Table 3.3: Composite media-based sentiment index's PCA results

	Per Cluster		Per Top 5	
	Eigenvalue	Percentage of Variance	Eigenvalue	Percentage of Variance
Comp 1	1.91	96.29	3.94	78.40
Comp 2	0.07	3.70	0.58	11.52
Comp 3			0.35	6.91
Comp 4			0.12	2.42
Comp 5			0.03	0.74

(a) Eigenvalues and Percentage of Variance from PCA analysis

Series (Contribution to dim)	Dim 1 (per cluster)	Dim 1 (per top 5)
Business Day Loughran	50	18.44
Sabinet Composite Loughran	50	
Business Day Bing		22.92
Business Day Harvard Four		17.78
News24 Bing		21.46
News24 Loughran		19.41

(b) Contribution of each constructed series to the two constructed MSI indices

3.5 Understanding the relationship between the CCI and the MSIs

To better understand the dynamic relationship between the media-based sentiment indices and the survey-based CCI we conducted two statistical tests. The first was to test whether the MSI and CCI series are cointegrated. If the results indicate that the series is indeed cointegrated, it provides evidence that a media based index could act as a possible alternative to the survey-based methods. Secondly, we test whether Granger-causality is present between the MSI and CCI. The result of this test would indicate whether the MSI could potentially lead the CCI and in doing so, acts as a complement. When testing for Granger-causality, the typical test can be problematic as using the F-statistic when one or both of the series are non-stationary can lead to spurious causality (He and Maekawa 2001). With this in mind, Toda and Yamamoto (1995)'s influential paper suggests an alternative framework where a VAR can be estimated in levels that allow for the testing of general restrictions on the parameter matrices, even if the underlying series are integrated or cointegrated. Figure 3.7 visually highlights the relationship over time between the constructed media-based indices (dotted) and the CCI (solid).

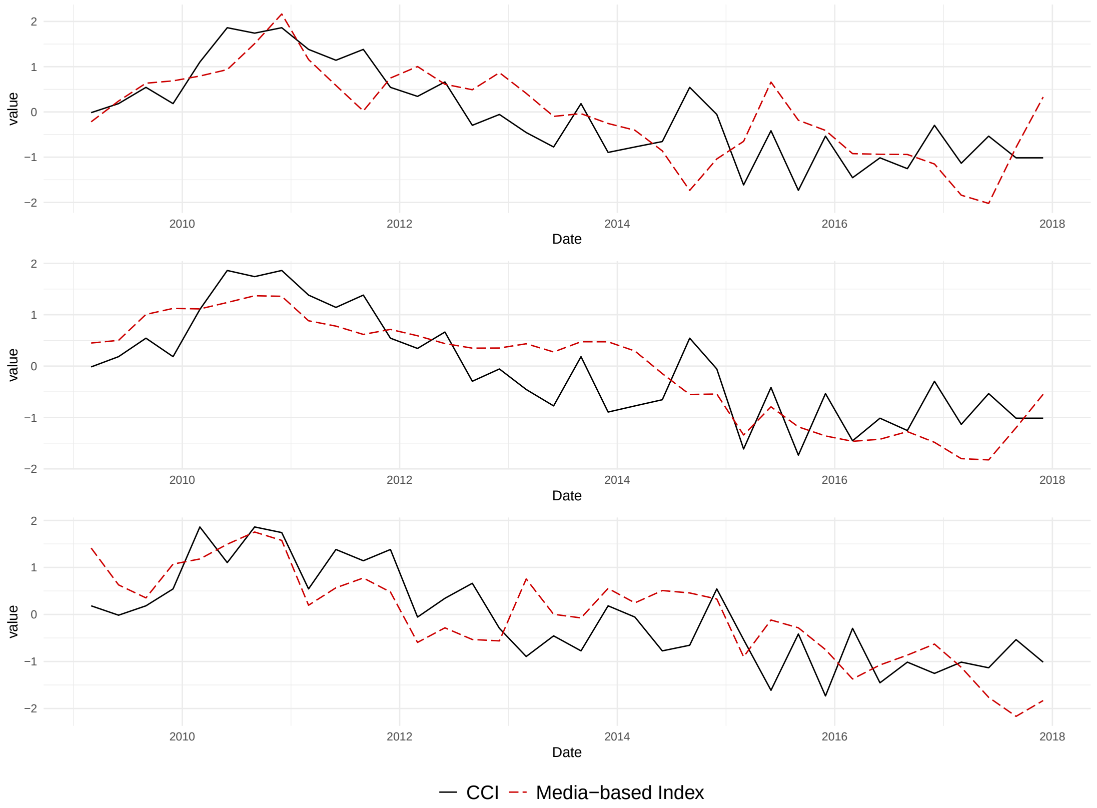


Figure 3.7: The figures show the relationship that the media-based sentiment indices have with the CCI after the quarterly transformation.

As with any VAR analysis, the series needed to be tested to see if they contain a unit root process and if they do, are they perhaps cointegrated? Given the small sample size, multiple unit root tests were conducted, namely ADF (Said and Dickey (1984)), Phillips-Perron (Phillips and Perron (1988)), and KPSS (Kwiatkowski et al. (1992)). For the ADF and KPSS test, all series were deemed to contain a unit root, while the Phillips-Perron suggest the series are stationary (see Appendix B.1). To test for cointegration, Johansen (1991)'s test was conducted and results showed that the MSI (Per Cluster) index and CCI is cointegrated, while the $H_0 =$ no cointegration, could not be rejected for the other two media based indicators (MSI per dtw, MSI news24). This result would indicate that MSI per cluster would be a good candidate as an alternative to the CCI, due to the series being integrated and thus does not deviate from equilibrium in the long term.

Next, we analysed the relationship between the indices using the Toda-Yamamoto (T-Y) procedure.[17] This procedure is used to estimate whether the media-based indices could be considered to Granger-cause the survey-based CCI.

Table 3.4: Toda-Yamamoto Granger-causality results

Granger causality H0:	Wald statistic	P-value
CCI does not Granger-cause MSI (Per Cluster)	5.57	0.06
MSI (Per Cluster) does not Granger-cause CCI	19.98	0.00
CCI does not Granger-cause MSI (Per DTW)	0.71	0.70
MSI (Per DTW) does not Granger-cause CCI	25.94	0.00
CCI does not Granger-cause News24 Loughran	0.70	0.71
News24 Loughran does not Granger-cause CCI	21.97	0.00

Table 3.4 shows that the MSI could be considered a good predictor of the CCI in all cases, while for the CCI, it can be said to only Granger-cause the MSI (Per Cluster). This supports the hypothesis that news media indices could potentially be capturing information on the sentiment of agents not yet reported in the official figures. To see if the media-based indices are not only useful in predicting the CCI, but lead it, cross-correlations were estimated. Due to being non-stationary, all series were differenced before the correlations were calculated.

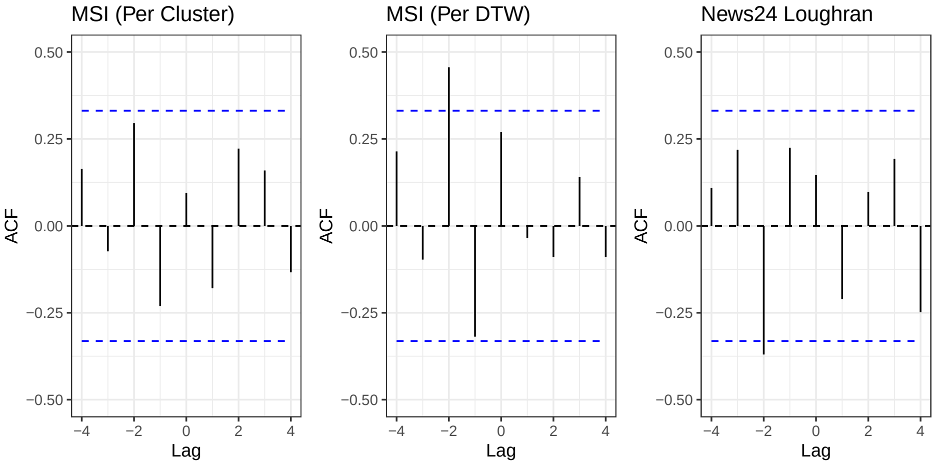


Figure 3.8: Cross-correlations between the MSIs and the CCI.

[17] The full 13-step sequence can be found at Toda and Yamamoto (1995).

Figure 3.8 shows that the MSI (Per DTW) and News Loughran indices could potentially be leading the CCI by two quarters as per the significant autocorrelation at lag $t - 2$.

3.6 Conclusion

The consumer confidence index is highly valued as a source of information used to forecast private consumption and commercial activity. The index contributes towards better understanding of economic business cycles, gives an indication of future economic activity, and provides insight into current economic conditions.

This chapters's aim was to investigate the feasibility of constructing online sentiment indices using various different text sources. We suggest the use of a clustering framework to select the most appropriate data sources and lexicon dictionaries to apply the bag-of-words approach. Our main objective was to investigate whether an online-based sentiment index can offer an complement to the survey-based approach that does not suffer from the same logistical challenges. If a mildly representative index is constructed, it has the advantage over the currently implemented BER confidence indices, which are only released quarterly and requires field surveys to be conducted. Emphasis is placed on how multiple indices, controlling for data provider and lexicon dictionary, can be tested as candidate alternatives for the traditional survey-based indices. This was done by employing a time-series clustering technique using dynamic time warping as a dissimilarity measure. Using the resulting clusters from the clustering as evidence for comovement, the series in the CCI cluster are used to create composite indices. Along with the cluster-based approach for a composite index, the raw dissimilarity values were also used to select the top n most similar series to the CCI. The results conclude that it is possible to create an index using sentiment analysis techniques that leverage large amounts of online editorial data that resembles the BER's consumer confidence index.

To investigate the relationship between the series, we test for cointegration as well as Granger-causality between the MSIs and the CCI. The results indicate that there exists a contemporaneous relationship between some of the constructed media-based indices and the survey-based consumer index, suggesting that a construct of a media-based sentiment exist that could act as an alternative measure of consumer confidence. In terms of the results from the Granger-causality analysis, all MSIs were deemed to be good predictors of the future values of the CCI. These findings not only support the use of media-based consumer confidence indices as an alternative indicator to measuring consumer confidence, but as a potential leading indicator to the survey-based index. This hypothesis is further corroborated by the cross-correlation results that indicate that the MSIs significantly led the CCI with two quarters.

To confirm the validity of these series as economic indicators, further research needs to be conducted on whether the series are able to predict future consumer and business activity out of sample. Another topic of interest would be to see whether the series provide informational content above and beyond the current measures of confidence in a forecasting experiment. Refinement in the construction of the indices (such as topic modelling) also need to be investigated, as this chapter aims only to suggest a generalised framework for the investigation of high-dimensional data to track economic sentiment.

In the next chapter these questions are explored when a media based sentiment index is constructed without the use of the clustering technique as a guidance. The index is then evaluated against the survey-based consumer confidence in forecasting personal consumption expenditure.

Chapter 4

Predictive power of text based sentiment indices

> *Take the output of machine learning algorithms (i.e. random forests) not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem. - Leo Breiman*

There is no refuting that there exists a contemporaneous correlation between sentiment and consumer spending, nor does it contradict the permanent-income models of consumption first pioneered by Friedman (1957) and empirically tested by Hall (1978). One of the earliest examples of the phenomenon where the contemporaneous relationship between sentiment and spending was observed is the start of the Gulf War. The index of consumer confidence fell 24.4 index points and what followed was an economic slowdown that followed (Carroll, Fuhrer, and Wilcox 1994). This drop in consumer confidence has since been cited as an important, if not the leading cause behind the contraction of household spending. The relationship between consumer confidence and consumer spending is again empirically shown to exist in Ludvigson (2004). The paper reports that most of the popular surveys of consumer confidence do contain information about the future path of the aggregate consumer expenditure growth.

Unfortunately for forecasters, forecasting changes in consumption remains difficult. From the point of an economic forecaster, the primary questions of interest are twofold: firstly, whether an index of consumer confidence has any predictive power on its own for future changes in consumption spending; secondly, whether the consumer confidence index contains information about future changes in consumer spending aside from the information contained in other macroeconomic variables available to the forecaster in the same period (Carroll, Fuhrer, and Wilcox 1994).

One of the factors inhibiting the construction and forecasting of national statistics or economic indicators constructed using surveys is the persistent pressure to reduce administration costs while maintaining a significantly representative sample. In this chapter, we investigate one such economic indicator: consumer confidence for South Africa. In this thesis, I construct quarterly media-based sentiment indices (MSIs) from large corpus data spanning from 2001 to 2017 using two well-known financial dictionaries. This chapter will use the terms sentiment and confidence interchangeably to refer to what is known as consumer confidence/sentiment. For the rest of the chapter, the collection of these indices will be referred to as media-based sentiment indices (MSIs). The predictive performance of the indices is compared to the traditional survey-based consumer confidence (CCI) in a Bayesian Vector Autoregressive (VAR) framework. This chapter's key focus is to explore the accuracy with which MSIs predict personal consumption expenditure (PCE) as well as its multiple sub-components.

Included in the forecast framework is a predictor of future PCE, disposable income[1], that acts as a baseline for forecasting accuracy. The inclusion of disposable income in the modelling process is to determine whether the MSI on its own has any predictive power above and beyond what is available to the forecaster in the same period. We compare the confidence indices by forecasting personal consumption expenditure and evaluating the out-of-sample performance through density-based scoring models as well as a statistical technique known as model confidence sets. The results show that the incorporation of text-based sentiment indices either improve out-of-sample forecasts for certain PCE measures, or perform at least as well as the existing survey-based CCI index. This result holds not only for short forecast horizons, but also for longer forecasting periods of up to two years.

[1] See Ludvigson and Steindel (1998) for theoretical basis.

In section 4.1 and 4.2, we review the relationship between consumer confidence and private consumption. We also consider why media sources could potentially be considered a good candidate for capturing overall consumer sentiment and personal consumption expenditure. Section 4.3 introduces the current literature on the subject of text-based sentiment indices with the aim of highlighting the difference between social media and online news text sources. We also explain how a media-based sentiment index can be developed using text.

Section 4.4 analyses the relationship between personal consumption expenditure, consumer confidence, and media-based sentiment. We are particularly interested in the correlation dynamics between the aforementioned series. This is followed in section 4.5 by the introduction of the econometric methodology used to evaluate the forecast performance of the different sentiment indices. Multiple Bayesian VARs are constructed and estimated using pseudo out-of-samples in an expanding window fashion from Q1 in 2007 to Q3 in 2017. The model estimates are evaluated using the density estimate of the forecasted value using scoring models and model confidence sets in section 4.6. The results also include a discussion of the resulting performance of the different indices and models during the global financial crisis.

4.1 Consumer confidence and consumer spending

In a seminal paper on the relationship between consumer confidence and the future path of consumer spending, Ludvigson (2004) found that consumer attitudes contributed significantly to explaining future consumption growth although the reason for this remains highly debated still. Traditional survey-based confidence indicators are typically used to provide a leading signal for real economic activity, especially with regard to personal consumption expenditure. Confidence in these indicators' ability to explain current and future economic developments is reflected in the extent of media exposure the CCI gets. In the US, well-known indices such as the Index of Consumer Sentiment (ICS) released by the University of Michigan, and the Consumer Confidence Index (CCI) issued monthly by the Conference Board, are widely integrated into the assessment of present and future economic paths of economic agents (Dominitz and Manski 2004). The information captured in these indices is believed to inform on household consumption decisions and the extent to which the agents of a household are willing to make new purchases and possible future spending behaviour. A consumer confidence index as such provides a key insight for analysts who are attempting to analyse the current and future business cycles of the economy. The CCI, although not always looked at in isolation, does play an important role in forecasting other macroeconomic variables, both financial and real. This is due to the index complementing information inherently contained in personal expenditure statistics.

Despite substantial exposure of consumer indices in the public domain, little consensus has been reached around the true predictive power of consumer confidences indices, especially after controlling for fundamental economic variables. The work of Bram and Ludvigson (1997) and Carroll, Fuhrer, and Wilcox (1994) provides evidence that lagged values of the CCI indicate that the index can improve short-term forecasts within the US. Other authors such as Acemoglu and Scott (1994) found similar results for the UK. Carroll, Fuhrer, and Wilcox (1994) found that using the ICS and its lags, the index explained around 14% of the variation in total real personal consumption expenditure for the period 1954 to 1994. The results were less convincing after controlling for other economic variables which would also have been available to the economic forecaster in the same time period. The economic variable Carroll, Fuhrer, and Wilcox (1994) used in their analysis to control for macroeconomic climate is real labour income, defined as wages and salaries plus transfers minus personal contributions for social insurance. Bram and Ludvigson (1997)'s paper specifically investigates the US example and compares the ICS with the CCI in having forecasting power of consumer spending. The authors found that the indices released by the Conference Board had both economically and statistically significant explanatory power for several categories of consumer spending. This was in contrast with the ICS released by the University of Michigan. The indices generally exhibited weaker forecasting power for most categories of spending. The research highlighted the key differences underlying the two surveys and how certain questions are more important than others in explaining the difference in predictive power. Fuhrer (1993) find that up to 70% of the University of Michigan Consumer Sentiment Index (ICS) can be explained by conditioning on other macroeconomic variables such as GNP growth, inflation rates, unemployment, and interest. Their research suggests that what the index captures is the reflection of general knowledge of the economic agents on the underlying economic climate. Even though the ICS measure appears to only have incremental predictive power once fundamentals are considered in the model framework, it does appear on a timely basis and as such plays a part in understanding consumption growth.

In other international literature, Gelper, Lemmens, and Croux (2007) used an approach whereby they decomposed the Granger causality at different time lags for the interaction between the sentiment index and PCE. Their results indicate that the US consumer sentiment index Granger causes future consumption with an average time lag of four to five months. Furthermore, their paper states that the sentiment index is better suited as a predictor for services than consumption of durables and non-durables. The forecast horizon is also of interest to economic forecasters. Although the CCI is generally considered to have predictive power for short-term forecasts such as $h = \{1, 2\}$ for quarterly horizons, Wilcox (2007) documented results that give evidence for the predictive power of confidence indices at longer horizons. The paper investigated the predictive power of the ICS and its individual questions at longer periods of up to four quarters ahead. The paper used as its baseline model variables that include consumption and its components, income and household wealth, as well as their lags. The results show not only that the inclusion of the ICS (and its separate questions) into the baseline model reduces error in the forecasts, but it does so for total expenditure, its components, and at the one-quarter-ahead and four-quarters-ahead horizons.

In contrast with the results of research in the US and UK, Fan and Wong (1998) found that for Hong Kong, various consumer confidence indices provide almost no explanatory power in forecasting consumption growth. Their paper argues that consumer confidence indices do not provide any new information above the fundamental economic and financial indicators such as labour income, real share prices, and short-term interest rates. Other fundamental variables such as personal disposable income also play a large role in explaining consumer spending as it affects spending behaviour through the wealth channel (Ludvigson and Steindel 1998). Their results posit that as the wealth of individuals increases, their propensity to consume increases as well.

In the context of South Africa, Khumalo (2014) investigated time-series data from 1980 to 2012 to test the existence of any possible long-run relationship between consumer spending and consumer confidence (as measured by the CCI). He employed a VAR model and used Granger causality to establish that there exists a relationship between consumer confidence and spending within South Africa.

4.2 Consumer sentiment, personal expenditure, and the media

Information absorbed by economic agents through the media is likely to have an impact on their attitude towards the current and future state of the economy. Barsky and Sims (2011) found that the information or 'news' channel has high explanatory power. By estimating a novel structural VAR approach that identifies news shocks about future technology, they found that a favourable news shock leads to an increase in consumption and decreases output, hours, and investment in the medium term. However, there seems to be little evidence of news contributing to the understanding of recessions.

In trying to better understand the relationship between informational channels (such as media), consumer sentiment, and consumer spending, we draw from the literature on information theory and decision making. The literature on sticky expectations and rational inattention emphasises the cost as well as the constraints in obtaining data and reacting to said data. Information theory provides a view on the role of media and the effect that the media has on influencing consumer sentiment on the economy. News media predominantly affects consumer sentiment through three different channels (Doms and Morin 2004). The first channel of information consists of the dissemination of economic statistics and opinions of experts over radio and printed media. Secondly, consumers receive a signal of economic activity through the tone and fluctuating volume of economic reporting. Over time, this signal may not always be consistent with real economic activity. Sims (2003) used the information theory model to show why tone and volume of economic reporting affect consumer sentiment more than economic information contained within the article. Headline articles conveying turning points in the business cycle, such as “Recession Possible”, elicit a greater negative response from economic agents in comparison to an article with a neutral headline talking about the same economic event. The final channel through which media, consumer sentiment, and economic activity are connected relates to the belief of sticky expectations and rational inattention, i.e., the likelihood that consumers will update their expectations after receiving information through the media.

Carroll (2003)‘s study on inflation expectations suggests that the likelihood that agents update their expectations increases as the volume of media coverage increases. This relates to the cost of acquiring information when it is abundant. As the volume of information increases, the cost of acquiring information decreases, and as such the cost of updating expectations decreases as well. This results in an increase in the frequency of expectations and expenditure fluctuations. Akerlof et al. (2000) argue that headlines such as “Recession Possible” are more likely to be read as it has a direct impact on agents’ financial position. Doms and Morin (2004) found evidence that consumer sentiment is affected through all three channels. They also found that expectations are less sticky during high-volume news and economic agents adjust their expectations more frequently. One of the ways that economists try to capture the expectations of the economic agents on the current economy and the future state of the economy is to use consumer confidence indices. This is further discussed in the following section.

4.3 Text-based applications in economics

The availability of information in the recent decade has seen drastic changes, both in volume and velocity. Two main drivers of this exponential growth of online information are social media and news publishing houses that publish content online. This information revolution has opened up the possibility of discovering new avenues for exploring economic questions. However, national statistical institutes have for the most part continued to rely on traditional data-gathering methods that use probability sampling in combination with design-based or model-assisted inference for the production of their officially released statistics (Van den Brakel et al. 2017).

The declining response rates to surveys over the last decade have made it difficult to maintain the representative sample. This concern in delivering reliable official statistics has stimulated the search for alternative sources of information, particularly online text-based data. Online sources could potentially be used to construct complementary statistical indicators. The potential is that so-called big data-based indices might be capable of enriching the current statistical products, improving the performance of current forecasting models, or even be disseminated as new statistical products within their own right. One of the many advantages of using big data as an alternative source of information is that it is generated as a by-product of processes rather than part of the statistical production pipeline of a national institution. Examples of data sets that have been employed as part of economic analysis include, but are not limited to: time and location of network activity from mobile phone companies, company reviews, social media messages from Twitter and Facebook, internet search behaviour from Google Trends, and, of course, media articles published online by media houses. One of these sources that has attracted a lot of interest over the last decade is text data. Text data can be analysed by either applying machine learning-based natural language algorithms, or through lexicon dictionaries. Both techniques have the ability to derive a text sentiment score which can be used in the construction of time-series-based indices. These appropriately named 'sentiment indices' are constructed using text-based analysis, a field within computational linguistics. Two obvious advantages of text-based measures of economic tracking are the coverage and cost aspects thereof. Primary research such as surveys is inherently expensive to conduct, and can also potentially be subject to small sample bias (Ludvigson 2004), which also restricts the frequency at which the statistic can be delivered.

Daas and Puts (2014) used data from a number of social media platforms such as Twitter, Facebook, and Linkedin, and also included Dutch messages produced on websites, forums, and blogs to derive an indicator of overall sentiment in the Dutch population. Their paper explores in depth the relationship between social media and consumer confidence and whether the social media index could act as an alternative to a national indicator such as the CCI. Their results show that there is indeed a relationship between the derived online social media sentiment and the traditional survey-based CCI measure for the Netherlands. Van den Brakel et al. (2017) further investigate the Dutch social media index using a multivariate structural time-series approach. In their paper, they estimate whether the inclusion of social media in the production of Dutch administrative statistics improves its forecasting accuracy. The paper also addresses the question of whether alternative data sources can be seen as complete substitutes for the traditional survey techniques. To do this, the authors estimate whether the two time series are cointegrated. They argue that a statistically significant outcome provides some evidence that, although the data sources are collected differently, they can be seen as having been generated by the same underlying evolutionary process. Thus, the SMI can be seen as a substitute for the more traditional survey approach used to construct the CCI. Following the modelling procedure in Harvey and Chung (2000), Van den Brakel et al. (2017)'s results indicate that the Dutch CCI and SMI do indeed follow a similar evolutionary process and as such can be seen as substitutes, albeit that the underlying data being measured is different. While Daas and Puts (2014) and Van den Brakel et al. (2017) both used social media as their main source of text data, others such as Fraiberger (2016), Larsen and Thorsrud (2015), and Shapiro, Sudhof, and Wilson (2018) used economic news media as their main corpus of text to conduct economic analysis.

Fraiberger (2016) utilised a combination of the Loughran and McDonald (2011)[2] and Young and Soroka (2012) dictionaries to construct a sentiment index from the full corpus of economic news articles produced by Reuters across 12 countries over a 25-year period. The paper found that the constructed index not only tracked GDP at country level, but contained information on future GDP growth which was not captured by consensus forecasts. Another approach is to disaggregate the news in such a manner as to separate news containing topics of information such as manufacturing, economic climate, or political news. This is typically done using a statistical technique known as topic modelling.

Larsen and Thorsrud (2015) employed topic modelling to deconstruct business news into various topics that newspapers report on. Their aim was to show that these topics have predictive power for key economic variables. They also evaluated how shocks to the economic news channel cause large and persistent economic fluctuations, a permanent increase in productivity, and are especially associated with financial markets, credit, and borrowing. The findings in their paper have

[2] For the rest of the paper this will be referred to as Loughran.

two major implications for the literature on text analysis in the economic context. Firstly, the results indicate that models where innovations in asset prices are used as a proxy for news shocks amplify effects of news and noise shocks. Secondly, decomposing business news into news topics could help explain how news shocks theoretically transmit and ultimately affect productivity and economic fluctuation. Text analysis of business news media can also be used to better understand how the channel of transmission between media, consumer sentiment and consumer spend functions.

Shapiro, Sudhof, and Wilson (2018) used online text data to try to capture economic activity. They propose the use of news media to develop a new time-series measure of economic sentiment based on both natural language processing and predefined lexicons. Their approach involved analysing economic and financial newspaper articles and constructing four alternative news sentiment indices by employing machine learning techniques that estimate sentiment. These sentiment indices were derived by evaluating the emotional valence of each article in terms of negativity, worry, and satisfaction. They also included a lexical “negativity” measure, resulting in four sentiment indices. The results show that news sentiment indices correlate strongly with contemporaneous business cycle indicators and that innovations in the index can outperform forecasts of economic activity. The paper illustrates how news-based sentiment indices can outperform the traditional Michigan and Conference board’s consumer confidence in predicting the federal funds rate, consumption, employment, inflation, industrial production, and the S&P500. The study concludes that news sentiment is a good contemporaneous and future indicator of economic outcomes even after conditioning on traditional survey-based alternatives.

4.3.1 Sentiment through textual analysis

Computational linguistics is best known for its sentiment and topic analysis toolset.[3] A body of text can typically be characterised by examining two facets within the text: (1) the degree to which the text exhibits emotion compared to a neutral

[3] To read up on how topic analysis is being applied within the field of economics and finance, see Hansen and McMahon (2016), Larsen and Thorsrud (2015), or Hansen, McMahon, and Prat (2014).

stance, and (2) the degree to which a certain emotion is deemed to be dominant in the writing. The psychology literature usually divides these emotions into two dimensions: valence and arousal. Valence captures the intrinsic goodness (positivity) or averseness (negativity) towards a subject, object or body of text; arousal describes and measures the intensity of the emotion.

A dictionary-based bag-of-words (BOW) approach to sentiment mining is widely used, partially due to the ease of use and also because the words tend to stay constant over time. The text is vectorised and sees a document collapsed down to a term-document matrix consisting of rows of words and columns of word counts based on predetermined word lists. This approach uses predefined dictionaries which consist of words associated with emotion. In other words, a "dictionary" is a tabulated collection of items, each with an associated attribute, for example (but not limited to), positive or negative. Each of the dictionaries has their own bag of associated words and can thus deliver very different results when estimating a sentiment score. Dictionary methods can also be used to capture uncertainty. Baker, Bloom, and Davis (2016), for instance, developed an economic policy uncertainty index using a lexical-based method that identifies and counts articles containing the word "uncertain" and "not certain" and combines these with terms related to economic policy. Using human-verified readings of the articles confirmed that the index proxies for movements in policy-related economic uncertainty.

In this chapter, we employ the well-known BOW technique to construct media-based indices from two established financial dictionaries: Loughran and McDonald (2011) and Henry (2008). The resulting word list of Henry (2008)‘s paper was the first lexicon that specifically tried to capture financial terms. She constructed the dictionary by examining earnings press releases from the telecommunications and computer services industries. One of the weaknesses of the Henry dictionary is the very limited number of words in the resulting lexicon. The dictionary only contains 85 negatively assigned words, while a dictionary like Loughran contains over 2300. In a study that examines how the tone of quarterly earnings conference calls relate to higher stock returns, Price et al. (2012) found that by employing the Henry dictionary, a significant market reaction to the press release affected both the initial reaction window and the 60-day drift period after the release. The study motivates that the Henry dictionary could be used to measure the tone of managers' reporting. In contrast with the Henry dictionary, the Loughran dictionary consists of six different word lists: negative, positive, uncertainty, litigious, strong modal, and weak modal. These lists were constructed by examining K-10 filings for the period 1994 to 2008. The sentiment lists are created based on the most likely interpretation of a word in a business context. The Loughran dictionary has two main advantages when compared to other financial dictionaries, for example, Henry. Firstly, the comprehensive list is aimed at making avoidance of key words more challenging, as well as ensuring no commonly appearing negative or positive words are missing. The second advantage of the dictionary is the fact that it was created with financial communication in mind, especially in media. Dougal et al. (2012) built on the familiar work of Tetlock (2007) by examining the "Abreast of the Market" column in the Wall Street Journal. They found that journalists associated with a more pessimistic column tone are directly linked to more negative market returns the following day. Various other examples on the application of financial dictionaries are available; we refer readers to Loughran and McDonald (2016) for an extensive discussion on lexicons.

4.3.2 Creating a sentiment score

The total articles, N_i^a, for a given data set $i \in \{BusinesDay_x, FinancialMail_x\}$ represent the two newspapers under investigation. The x term represents the individual articles within each data set, while the total period over which the analysis will be conducted is represented by $T^d = \{$2001-Q1:2017-Q3$\}$. As is standard with any analysis conducted on a corpus of this size, a data-preparation step was introduced before analysis could begin.

All data cleaning was done using R Core Team (2013) and the `tidytext` library by Silge and Robinson (2016).[4] The first step of cleaning raw text data is to remove all stop words.[5] Stop words are commonly used words such as *the*, *and*, *a*, etc. that do not contribute anything towards understanding the content of the underlying text. The lexicon we employed to remove the stop words contains 1149 stop words. A common second step of text analysis involves stemming. This leaves the core

[4] To clean the data, the tidyverse R package was used (Wickham 2017).

[5] The Bouchet-Valat (2019) library in R was used.

part of the word that is common to all of its inflections. For this thesis, we did not stem words, as the dictionaries that we used do not require this. The last bit of cleaning involves removing all editorial pieces with less than 150 words in order to have enough information to create a robust sentiment score.

To construct a sentiment score, we identified the positive and negative words in each article $N_i^a \quad \forall \quad T$ using an external word list (dictionary). Next, we vectorised the text and only selected the words that are both in the dictionary and the piece of text. We then used the dictionary to assign positive and negative sentiment to the mutual words as per the relative lexicon. Conducting a simple word count that consists of the positive plus negative words, we normalised the count so that it reflects the relative proportion of positive and negative words within an article:

$$\begin{aligned} Pos_{i,t,n^a} &= \frac{PositiveWords}{PositiveWords + NegativeWords} \\ Neg_{i,t,n^a} &= \frac{NegativeWords}{PositiveWords + NegativeWords} \end{aligned} \tag{4.1}$$

The overall sentiment score for each article n_i^a, for $n_i^a = 1, \ldots, N_t^a$ at day t can then be defined as:

$$S_{i,t,n^a} = Pos_{i,t,n^a} - Neg_{i,t,n^a} \tag{4.2}$$

The polarity of the article is derived from the score. If the score of the article is greater than zero, the overall sentiment for the article is deemed to be positive, and vice versa for a negative sentiment score. The index is constructed as the net balance of positive and negative articles within a quarter.

This method resulted in two media-based sentiment indices available to analyse, namely Loughran and Henry sentiment indices. We also constructed a third index. This index is a combination between the two indices. The weights for each of the indices in the composite index are derived by applying principal component analysis (PCA). The PCA analysis resulted in the composite index being created by weighing the Loughran index and the Henry index equally.

Unlike the smoothing adjustments made in Chapter 3, the sentiment indices constructed in this chapter undergoes no transformation or filtering. The reason behind this choice is the extended period of the analysis and because the sentiment is aggregated at a quarter and not monthly as in the previous chapter.

4.3.3 Data

The data source for the analysis comprises the Business Day and Financial Mail data set as described in section 2.2. The total amount of articles used in the analysis amounts to approximately 150 000.

In terms of the confidence indicator, a noticeable characteristic of the CCI's release date is that the release is not always on the first day of the quarter, e.g., 1 December (Q4) or 1 March (Q1). For example, the 2001 Q4 release of the CCI number occurred on the 5th of December, while 2002 Q1 figures were only released on the 11th of April.[6] Another example would be the 2016 Q3 and Q4 publications. In this case, the figures for Q3 were only released on the 7th of December, while the Q4 data was quickly available afterwards on the 18th of January 2017.[7] Due to media data having a daily frequency, the difference between days within a quarter was taken into account when we constructed the media-based sentiment indices. To compare the two different frequencies, each quarter is defined as the days between the quarterly release dates for the *PCE measure*. The alignment was carried out so as to ensure that the media sentiment index would reflect the amount of information available to the forecaster at the time of estimation the PCE measures were released, as opposed to the release dates of the traditional CCI.

4.4 Analysing the relationship between PCE, CCI, and MSI

To compare the survey-based CCI and the MSI in a forecasting framework, various personal consumption data was sourced from the South Africa Reserve Bank's

[6] This amounts to 128 days passing between the releases.

[7] In between the release of the figures, only 43 days passed.

data portal.[8] The data consists of personal consumption expenditure figures. The quarterly time series includes total, durable, non-durable, semi-durable, and services expenditure in South Africa over the period 2001 Q1 to 2017 Q3. Total PCE is a combination of all the expenditure indicators. The analysis also includes disposable income that aims to condition for macroeconomic fundamentals within the forecasting model. If the time series is non-stationary, the growth rate is taken using $ln(Yt/Y_{t-1})$, thus resulting in the data from 2001 Q2. Following Carriero, Clark, and Marcellino (2015), all time series were persistent and required transformation.[9]

Table 4.1: Data description, related SARB code, and the unit of measure before transformation.

Name	Code	Unit.of.Measure	Description
PCE tot	KBP6007D/L	R millions	Final consumption expenditure by households: Total (PCE)
PCE durable	KBP6050D/L	R millions	Final consumption expenditure by households: Durable goods (PCE)
PCE non durable	KBP6061D/L	R millions	Final consumption expenditure by households: Non-durable goods (PCE)
PCE semi durable	KBP6055D/L	R millions	Final consumption expenditure by households: Semi-durable goods (PCE)
PCE services	KBP6068D/L	R millions	Final consumption expenditure by households: Services (PCE)
Disposable income	KBP6246L	R millions	Disposable income of households

Figure 4.1 displays the relationship between the various media-based sentiment indices and the traditional survey-based CCI from the BER.

[8] https://www.resbank.co.za/Research/Statistics/Pages/OnlineDownloadFacility.aspx.

[9] The work of Carriero, Clark, and Marcellino (2015) on modelling in levels or growth rate showcases that growth rates have the potential to provide better forecasts. For more examples of work that uses growth rates in BVAR forecasting, see Del Negro and Schorfheide (2004), Clark and McCracken (2008), and Koop (2013).

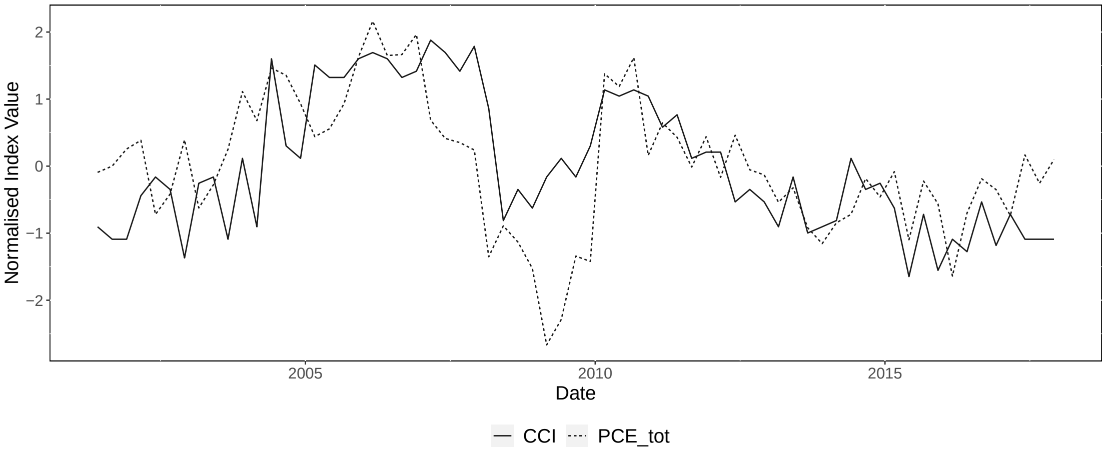


Figure 4.1: Comparison of survey-based CCI with total personal consumption expenditure.

Alongside the traditional survey-based CCI, our newly derived news measure is also under investigation. Figure 4.2 displays the relationship between the various media-based sentiment indices and the traditional survey-based CCI from the BER.

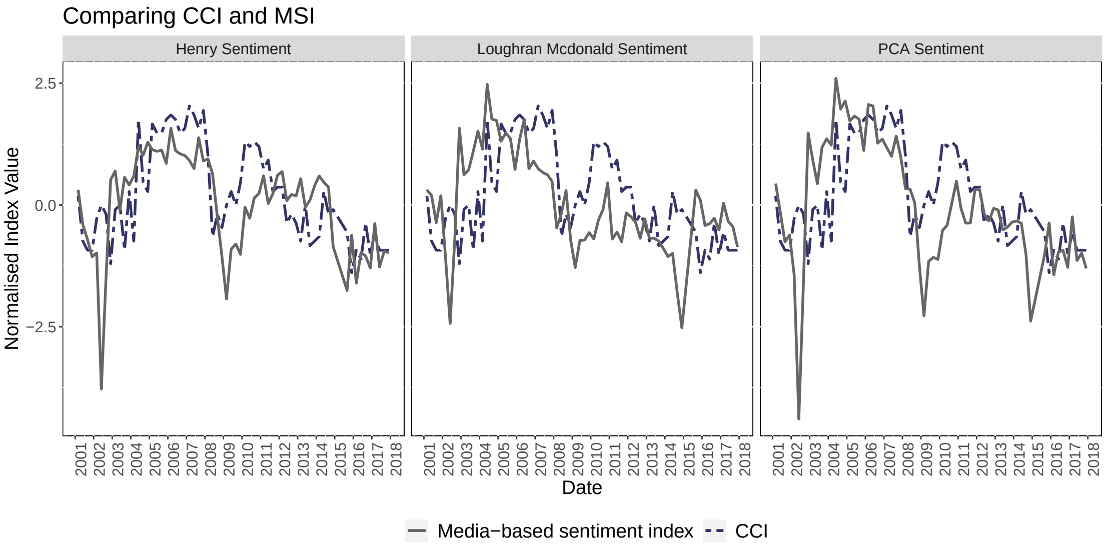


Figure 4.2: Comparison of media-based sentiment index (MSI) with the survey-based CCI.

The figures show how the choice of dictionary can be highly influential on the final index. The final observation of the individual dictionaries even display opposite directional behaviour. Figure 4.3 contains a comparison between the various media-based indices and the outcome variables of interest: personal consumption expenditure (PCE) indicators.

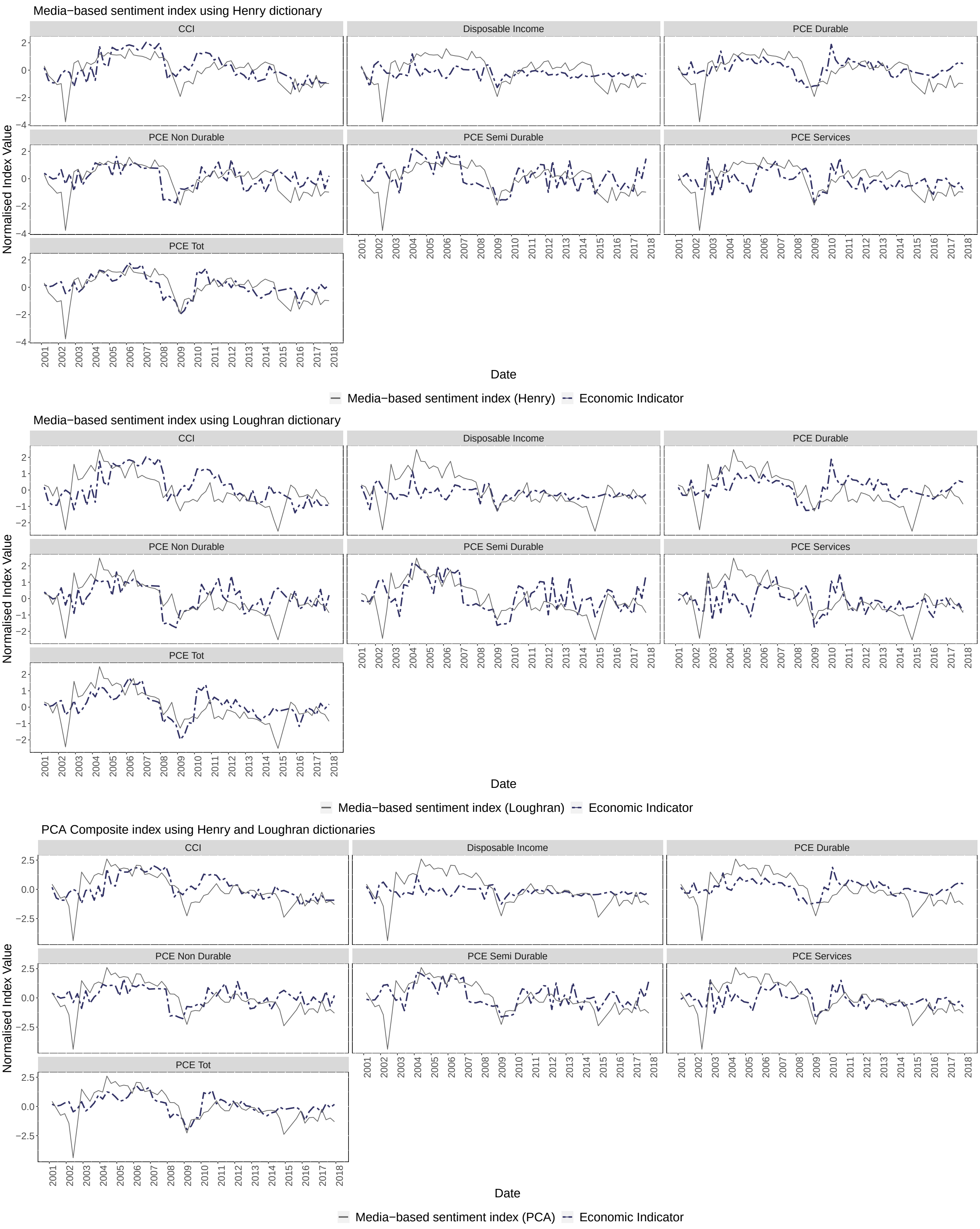


Figure 4.3: Comparison of media-based sentiment index (MSI) with various PCE sub-components as well as the survey-based CCI and Disposable income. PCA Composite index constructed from indices using Henry and Loughran dictionaries.

Figures 4.1 to 4.3 all show the co-movement between consumer sentiment, media sentiment, and log differenced personal consumption expenditure indicators. To see the strength of the comovement between the variables, we constructed a Pearson correlation coefficient between the sentiment indicators and the economic variables. These correlations are graphically communicated in figure 4.4. We also analysed the correlation of the CCI and the media-based indices based on lagged terms of the sentiment indices to evaluate whether the confidence indicators can be considered to have any leading properties.[10] (figure 4.5)

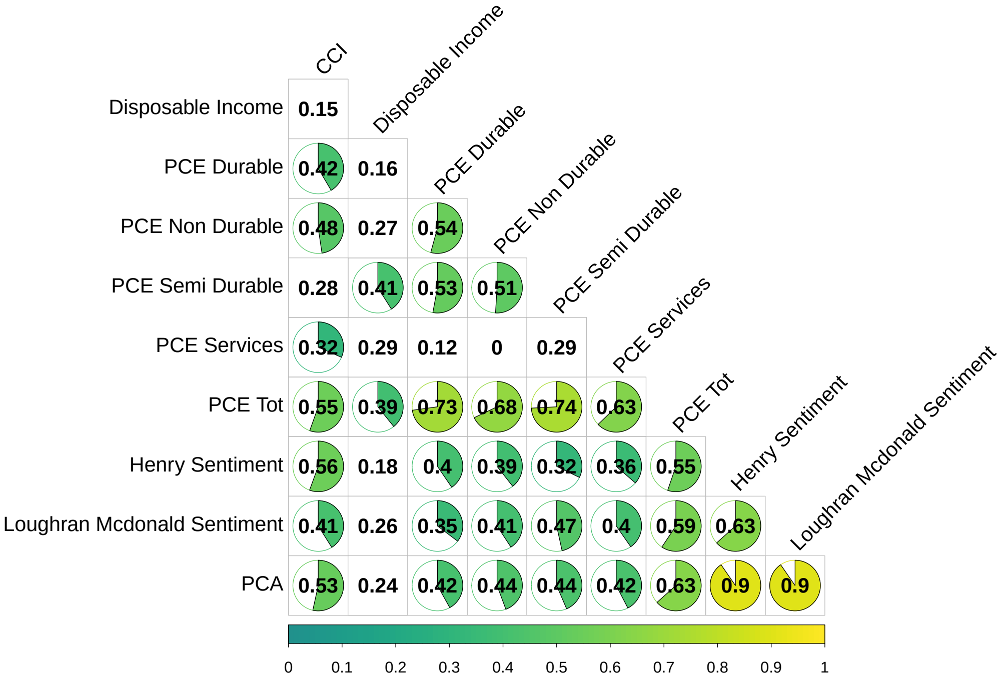
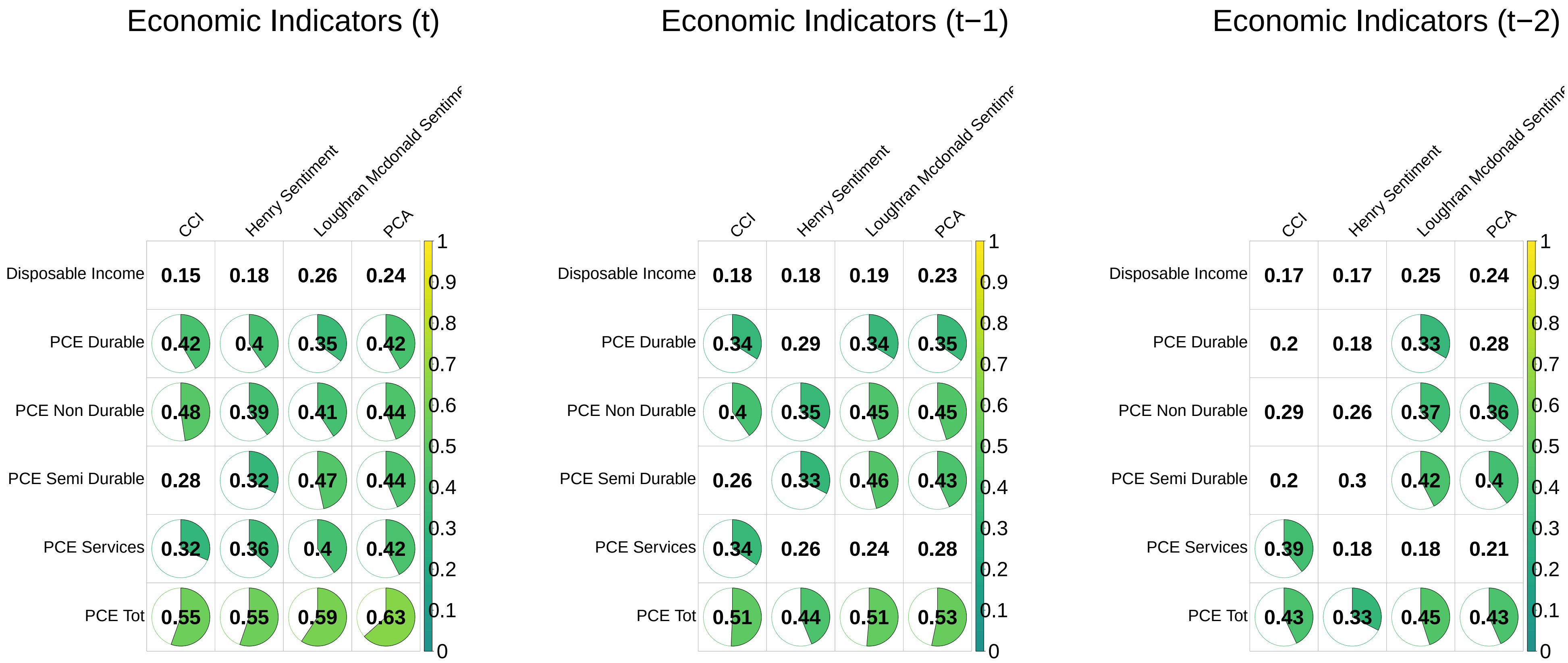


Figure 4.4: Measuring the co-movement between the various PCE and the MSIs using Pearson correlation.

[10] If a statistically significant relationship exists between two variables, the block contains a small pie chart.



Economic Indicators (t)

	CCI	Henry Sentiment	Loughran Mcdonald Sentime	PCA
Disposable Income	0.15	0.18	0.26	0.24
PCE Durable	0.42	0.4	0.35	0.42
PCE Non Durable	0.48	0.39	0.41	0.44
PCE Semi Durable	0.28	0.32	0.47	0.44
PCE Services	0.32	0.36	0.4	0.42
PCE Tot	0.55	0.55	0.59	0.63

Economic Indicators (t−1)

	CCI	Henry Sentiment	Loughran Mcdonald Sentime	PCA
Disposable Income	0.18	0.18	0.19	0.23
PCE Durable	0.34	0.29	0.34	0.35
PCE Non Durable	0.4	0.35	0.45	0.45
PCE Semi Durable	0.26	0.33	0.46	0.43
PCE Services	0.34	0.26	0.24	0.28
PCE Tot	0.51	0.44	0.51	0.53

Economic Indicators (t−2)

	CCI	Henry Sentiment	Loughran Mcdonald Sentime	PCA
Disposable Income	0.17	0.17	0.25	0.24
PCE Durable	0.2	0.18	0.33	0.28
PCE Non Durable	0.29	0.26	0.37	0.36
PCE Semi Durable	0.2	0.3	0.42	0.4
PCE Services	0.39	0.18	0.18	0.21
PCE Tot	0.43	0.33	0.45	0.43

Figure 4.5: Comparison of MSI with various PCE measures as well as the survey-based CCI and Disposable income. PCA Composite index was constructed from indices using Henry and Loughran dictionaries.

Figures 4.4 and 4.5 graphically illustrate the correlations between the consumer sentiment indices and the various PCE sub-components indicators. The figures also include whether the correlation coefficient between the relevant series are statistically significant or not. If the square with the respective correlation coefficient does not contain a small pie chart, the correlation is deemed not to be significant at a 1% level. This insignificance can be seen in the relationship between the PCE service component, the other PCE sub-components, and the consumer sentiment indices. We also see that although disposable income is considered a very good predictor variable in forecasting personal consumption expenditure, the contemporaneous correlation is non-significant. The CCI has the highest correlation with non-durables, Services, PCE total, the Henry media sentiment index, and the PCA media-based sentiment index. This relationship holds even when we lag the indicator, $t-1$. At $t-2$, the CCI still has a correlation coefficient of 0.51 with total PCE measure which is significant at 1%. The figures also show that the correlation coefficient between PCE services and CCI increases as the confidence indicator is lagged, increasing from 0.30 to 0.47.

For the MSIs, the indices have the highest correlation with total PCE as well as the service component. The Loughran and composite media-based sentiment indices have a high contemporaneous correlation with services, which quickly decreases as we lag the sentiment indices. With regard to total PCE, non-durables, and semi-durables, the media-based sentiment indices still have a significant relationship at $t-2$. The figures provide initial evidence that the traditional survey-based consumer confidence index and the media-based sentiment indices could assist in forecasting personal consumption expenditure and its sub-components. The correlation coefficients also indicate that the traditional and media-based indices could possibly provide complementary information as CCI correlates with durables and services, while the MSIs correlate with non-durables, semi-durables, and total PCE. Next, we formally tested the relationship between the sentiment indices and economic indicators within an econometric framework by estimating multiple Bayesian Vector Autoregressions (BVARs) at different forecast horizons and measuring their performance at each horizon. We conducted the analysis at multiple forecast horizons so as to investigate the robustness of the results both in the short and long run.

The choice of using a BVAR model, as opposed to a simple time-series regression, stem from the considerable use of these model in policy analysis and forecasting macroeconomic variables (Carriero, Clark, and Marcellino 2015). BVAR models have also shown to produce highly accurate forecasts and provides information on the confidence of the mean point forecast - useful in communication of forecasts of indicators such as inflation and GDP growth. In the case of this thesis, the hypothesis is that the inclusion of sentiment indices should not only improve the accuracy of the forecast, but also decrease the uncertainty around it and Bayesian techniques allows us to test this hypothesis.

4.5 Evaluating forecasting performance

Take note that this chapter performs the forecasting pseudo "real-time". This entails never using information that was not available at the time of forecasting and as such many researcher regard out-of-sample performance as the "ultimate test of a forecasting model" (Stock and Watson 2007). The forecasting is conducted using an expanding estimation from 2007 Q1 to 2017 Q3 resulting in 43 out-of-sample quarters (or test samples) to evaluate. This period also allows for the out-of-sample to contain the crisis period in the forecast evaluation. The forecasting model will produce forecasts for horizons 1, 2, 4, 6, and 8 quarters ahead at each of the 43 out-of-sample periods. Table 4.1 shows the different personal consumption expenditure indicators included in the forecast evaluation. The table also gives the respective SARB code for each expenditure indicator as per the Bank's bulletin. Alongside the expenditure indicators, we also include disposable income in the analysis so as to control for macroeconomic fundamentals. Figure 4.6 represents a graphical representation of the forecast design. The tree represents the BVARs that are estimated for each of the models constructed, as defined later in equation (4.13) to (4.16), with forecast horizons $h = 1, 2, ...8$, for each t in period 2007 Q1 to 2017 Q3.

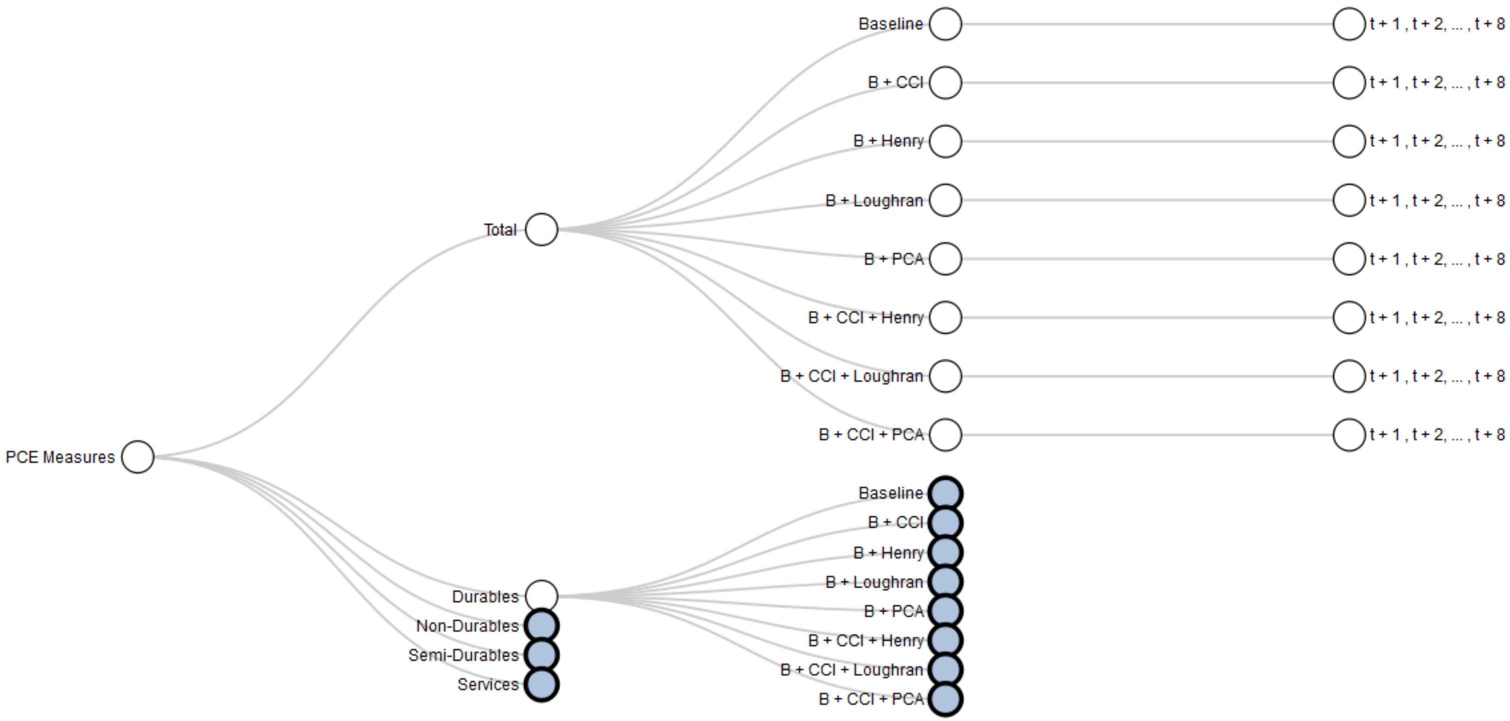


Figure 4.6: Tree representation of the forecast design where each of the leaves in the tree represents a BVAR forecast estimation.

4.5.1 Bayesian VAR

When selecting a specification for a Bayesian VAR, multiple construction and estimation nethods are available. Some of these methods can be technically and computationally demanding and such careful consideration needs to be taken before estimating such a model. In light of this, and following the evidence from Carriero, Clark, and Marcellino (2015), we specify the Bayesian VAR with a Normal-inverted Wishart (N-IW) conjugate prior with N-dimensional variables in a vector $y_t = (y_{1t}, y_{2t} \cdots y_{Nt})'$ with the VAR notation:

$$y_t = \Phi_c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \epsilon_t; \quad \epsilon \sim i.i.d.N(0, \Sigma) \tag{4.3}$$

where $t = 1, \ldots, T$. This results in each equation having $M = Np + 1$ regressors. This specification was empirically shown to provide negligible differences in accuracy when compared to more complex methods of estimation such as Markov Chain Monte Carlo (MCMC) simulation. If we group the coefficient matrices in the $N \times M$ matrix $\Phi = [\Phi_c \Phi_1 \ldots \Phi_p]$ and define $x_t = (1 y'_{t-1} \ldots y'_{t-p})'$ as a vector containing p lags of y_t along with a constant, the VAR can be written as:

$$y_t = \Phi x_t + \epsilon_t \tag{4.4}$$

which can also be written in a compact way as

$$Y = X\Phi + E \tag{4.5}$$

where the vectors represented by $Y = [y_1, \ldots, y_T]'$, $X = [x_1, \ldots, x_T]'$, and $E = [\epsilon_1, \ldots, \epsilon_T]'$ are respectively $T \times N$, $T \times M$, and $T \times N$ matrices. This chapter uses the conjugate N-IW prior:

$$\Phi | \Sigma \sim N(\Phi_0, \Sigma \otimes \Omega_0), \quad \Sigma \sim IW(S_0, \nu_0) \tag{4.6}$$

Due to the prior being conjugate, the conditional posterior distribution of the BVAR is also N-IW (Zellner 1971).

$$\Phi|\Sigma, Y \sim N(\bar{\Phi}, \Sigma \otimes \bar{\Omega}), \quad \Sigma \sim IW(\bar{S}, \bar{\nu}) \tag{4.7}$$

As in the work of Carriero, Clark, and Marcellino (2015), if we define $\hat{\Phi}$ and $\hat{E}$ as the OLS estimates, $\bar{\Phi} = (\Omega_0^{-1}X'X)^{-1}(\Omega_0^{-1}\Phi_0 + X'Y))$, $\bar{\Omega} = (\Omega_0^{-1}X'X)^{-1}$, $\bar{\nu} = \nu_0 + T$ and $\bar{S} = \hat{\Phi}'X'X\hat{\Phi} + \Phi_0'\Omega_0^{-1}\Phi_0 + \Phi_0 + \hat{E}'\hat{E} - \hat{\Phi}'\bar{\Omega}^{-1}\hat{\Phi}$. To obtain the 1-step ahead forecast $\hat{y}_{t+1}$ we can use the posterior mean $\bar{\Phi}$:

$$\hat{y}_{t+1} = \bar{\Phi}_c + \bar{\Phi}_1 y_t + \bar{\Phi}_2 y_{t-1} + \cdots + \bar{\Phi}_p y_{t-p+1} \tag{4.8}$$

The h-step ahead forecast is obtained by iteration:

$$\hat{y}_{t+h} = \bar{\Phi}_c + \bar{\Phi}_1 \hat{y}_{t+h-1} + \bar{\Phi}_2 \hat{y}_{t+h-2} + \cdots + \bar{\Phi}_p \hat{y}_{t+h-p} \tag{4.9}$$

where $\hat{y}_{t+h} = y_{t+h-p}$ for $h \leq p$. This iterative formulation can also be represented in a compact notation as in equation (4.4):

$$\hat{y}_{t+h} = \Phi^h x_t \tag{4.10}$$

The standard model implemented in this chapter will impose prior expectation and variance coefficient matrices as:

$$E[\Phi_k^{(ij)}] = \begin{cases} \Phi^*, & \text{if } i = j, k = 1 \\ 0, & \text{otherwise} \end{cases}, \quad SD[\Phi_k^{(ij)}] = \begin{cases} \frac{\lambda_1 \lambda_2}{k} \frac{\sigma_i}{\sigma_j}, & k = 1, \ldots, p \\ \lambda_0 \sigma_i, & k = 0 \end{cases} \tag{4.11}$$

where $\Phi_k^{(ij)}$ denotes the element in position (i, j) in the prior matrix Φ_k. The covariances in the Φ_k matrix are set to 0, while for the intercept the chapter assumes an informative prior with mean 0 and standard deviation $\lambda_0 \sigma_i$.

The tightness of the prior is measured by the shrinkage parameter λ_1: when $\lambda_1 \to 0$. This is the case when the prior is imposed exactly and the data did not influence the estimate. The same holds for when $\lambda_1 \to \infty$ the prior becomes loose and the prior information does not influence the estimates, which will approach the standard OLS estimates. The model also contains another shrinkage parameter, λ_2, which implements additional shrinkage on the lags of the other variables instead of the lags of the dependent variable. Carriero, Clark, and Marcellino (2015) refers to this parameter as the cross-shrinkage parameter, and as is the case in their model, we set it to $\lambda_2 = 1$. This specification of λ_2 implies no cross-shrinkage is applied as is needed for the N-IW specification. In specifying the priors, we follow Carriero, Clark, and Marcellino (2015) who shows that good priors for economic applications are:

$$\lambda_0 = 1; \quad \lambda_1 = 0.2; \quad \lambda_2 = 1; \quad \lambda_3 = 1; \quad \lambda_4 = 1 \tag{4.12}$$

Where λ_3 and λ_4 aim to specify the autoregressive and stationarity of the variables, respectively. For a full mathematical explanation of the model constructs and how the Bayesian VAR and its prior parameterisation are formalised, we refer readers to Carriero, Clark, and Marcellino (2015).

Equations (4.13) to (4.16) illustrate how we construct the baseline along with the extensions to the baseline which include the traditional survey-based CCI and/or MSI.

$$\text{Baseline:} \quad y_{i,t+h} = \sum_{l=0}^{4} Y_{t-l} + \sum_{l=0}^{4} DI_{t-l} + \varepsilon_{i,t} \tag{4.13}$$

$$\text{Baseline + CCI:} \quad y_{i,t+h} = \sum_{l=0}^{4} Y_{t-l} + \sum_{l=0}^{4} DI_{t-l} + \sum_{l=0}^{4} CCI_{t-l} + \varepsilon_{i,t} \tag{4.14}$$

$$\text{Baseline + News Sentiment:} \quad y_{i,t+h} = \sum_{l=0}^{4} Y_{t-l} + \sum_{l=0}^{4} DI_{t-l} + \sum_{l=0}^{4} S_{i,t-l} + \varepsilon_{i,t} \tag{4.15}$$

$$\text{Baseline + CCI +News Sentiment:} \quad y_{i,t+h} = \sum_{l=0}^{4} Y_{t-l} + \sum_{l=0}^{4} DI_{t-l} + \sum_{l=0}^{4} CCI_{t-l} + \tag{4.16}$$

$$\sum_{l=0}^{4} S_{i,t-l} + \varepsilon_{i,t}$$

That is, for each forecast horison $h = \{1, 2, 3, 4, 6, 8\}$, a VAR regression with 4 lags of the PCE component and disposable income (DI), represented by Y_i, is estimated in a Bayesian framework. If a sentiment measure, CCI or S_i, is included, 4 lags of the sentiment is also included. The first two extended models, equation 4.14 and 4.15, are the standard horse-race forecasting competition. That is to say, we aim to determine whether sentiment indices, survey- or media-based, have any predictive power independently of one another. Equation 4.16's aim is to determine whether the inclusion of a news sentiment index provides any additional forecasting information over and above the traditional survey-based CCI which is also available to the forecaster.

To evaluate the model forecast, we used density-based evaluation methods. Traditionally, deviations from the true realisation are considered to represent forecast performance. Metrics such as RMSFE, MAFE, and MFE are used most often, but can be problematic when the forecast distribution departs from normality. Failing to take account of uncertainty in the forecast can lead to biased conclusions of the model's forecasting ability. This is especially true in the case of policy makers who are explicitly concerned with the forecast interval or density-based forecasts (commonly known as fan charts) when it comes to policy design.

4.5.2 Density-based error measures

In order to evaluate marginal density out-of-sample forecasts with a point realisation, density-based error measures or *scores* were employed. The most popular scoring measures are the Log Score (LS) and Continuous Ranked Probability Score (CRPS). These two specific scoring measures are also considered to be *strictly proper* scoring rules and form attrative summary measures of predictive performance in that they address both calibration and sharpness[11] simultaneously (Gneiting, Balabdaoui, and Raftery 2007). Another favourable property of the CRPS is that it generalizes the mean absolute error. Therefore it provides a direct way of comparing deterministic and probabilistic forecasts using a single metric, while also being expressed in the same unit as the observed variable. For a technical discussion on scoring rules, see Gneiting and Raftery (2007) and, more recently, Krueger et al. (2016). Following Gneiting and Ranjan (2011), we define log scores as negatively oriented penalties represented as:

$$LS = -log(p_i(x^0)) \tag{4.17}$$

where $p_i(x^0)$ is the value of the predictive density of variable X at the realisation x^0. The LS scoring method is easy to calculate, but can be sensitive to change in small samples, more so if the choice of prior in the Monte Carlo simulation changes. Thus, the model is better suited to financial applications where the number of observations is much greater, as is the case in Weigend and Shi (2000). This also has relevance to low probability events, as the effect of the logarithmic transformation can impact evaluations where outlier events are important to capture. Given the shortcomings of the LS method, an alternative density-scoring method CRPS is also estimated. For CRPS, the empirical CDF is defined with the use of an indicator function $\mathbb{1}_{\{x^0 \leq x\}}$:

[11] Following Gneiting, Balabdaoui, and Raftery (2007): Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only.

$$CRPS(F, x^0) = \int_{-\infty}^{+\infty} [F(X) - \mathbb{1}_{\{x^0 \leq x\}}]^2 dx, \tag{4.18}$$

where $\mathbb{1}_{\{x^0 \leq x\}}$ states that if $x^0 \leq x$, then it is 1, otherwise it is 0. If the first moment of F is considered to be finite, the CRPS is represented in terms of the predictive CDF F is given by:

$$CRPS(F, x^0) = E_F|X - x^0| - \frac{1}{2}E_{F,F}|X - X'|, \tag{4.19}$$

where X and X' are considered to be i.i.d random variables from the same distribution F. Unfortunately, in applications such as Bayesian estimation, the forecast distribution of interest F is not available in an analytical form, but is formed from simulated samples $X_1, \ldots, X_m \sim F$. These simulated samples are converted to a distribution with a closed-formed expression. We estimate the CRPS using Jordan, Krueger, and Lerch (2017)'s package in R which is based on a numerically efficient estimation of the generalised quantile function:

$$CRPS(\hat{F}_m, x^0) = \frac{2}{m^2} \sum_{i=1}^{m} (X_{(i)} - x^0) \left(m\mathbb{1}\{x^0 < X_{(1)}\} - i + \frac{1}{2} \right) \tag{4.20}$$

In order to estimate the approximation $\hat{F}_m$, we use the empirical distribution function as per Laio and Tamea (2007)'s version of the quantile decomposition of the CRPS.

For both CRPS and LS, the scoring method has a symmetric loss function interpretation, but CRPS is more symmetric around the minimum and, as such, rewards a better score to forecasts closer to the realisations of the forecast density. Gneiting and Raftery (2007) also reports that CRPS is less sensitive to outliers.

To evaluate the forecasts at different horizons, summary statistics of the two density-based scores are considered. These statistics include mean, median, and standard deviation. Given the distributional properties of the scores[12], the median value is considered to be the best suited for evaluating a model's forecasting performance. We also include a traditional point error metric, root mean squared error (RMSE), for comparison (Chai and Draxler 2014).

[12] The distribution of the scores tended to be heavy right-tailed and not a normal distribution.

4.5.3 Model Confidence Sets

In the case where there exists multiple model specifications, all adequate in describing the underlying data-generating process, the question of "best model" seems ambiguous. In order to better validate what is considered to be the optimal model, P. R. Hansen (2005) proposes a statistical procedure that consists of a sequence of statistical tests that allow for the construction of what is known as a "Superior Set of Models" (SSM) from Model Confidence Sets (MCS). The SSM is established when the null hypothesis of equal predictive ability (EPA) can no longer be rejected at a user-specified confidence level. What makes this test versatile is the fact that the EPA test statistic is loss function agnostic, which permits the testing of models on various aspects, e.g., in-sample forecasts as demonstrated in Hansen, Lunde, and Nason (2011), or point-forecasts observed in P. R. Hansen and Lunde (2005). The advantage of the MCS procedure over a more well-known goodness of fit test, such as Diebold-Mariano, is the number of tests that need to be performed. For n models, we would have to perform $n(n-1)/2$ hypothesis tests to get all the pairwise comparisons between the model sets. In comparison, the MCS test only needs to be performed once. In addition, the Diebold-Mariano test is also not suited towards evaluating nested models (Diebold 2015).

Formally, let Y_t represent the realisation of the observation at time t and $\hat{Y}_{i,t}$ the forecasted value for model i at time t. A generic loss function, $l_{i,t}$ associated with the i-the model is then defined as

$$l_{i,t} = l(Y_t, \hat{Y}_{i,t}) \tag{4.21}$$

and represents some difference between the output $\hat{Y}_{i,t}$ and the *a posteriori* realisation Y_t. As mentioned, the loss function is arbitrary in the estimation procedure and will depend on the nature of the problem and the scope of the evaluation. The loss function is defined by the CRPS, LS, and squared error as formulated in 4.5.2.[13] From Bernardi and Catania (2018), the procedure starts from an initial set of models $\hat{M}^0$ of dimension m that consists of all the models under consideration. The output of the procedure is a smaller set, the superior set of models $\hat{M}^*_{1-\alpha}$ of

[13] We refer readers to P. R. Hansen and Lunde (2005), Bollerslev, Engle, and Nelson (1994), Diebold and Lopez (1996), and Lopez (2001) for different measures of loss functions.

dimension $m^* < m$, given a confidence level $1-\alpha$. The best scenario outcome from the procedure is when $m^* = 1$, representing a final set of a single model. Next, let $d_{ij,t}$ represent the loss differential between models i and j:

$$d_{ij,t} = l_{i,t} - l_{j,t}, \qquad i,j = 1,\ldots,m; \qquad t = 1,\ldots,n \tag{4.22}$$

while

$$d_{i\cdot,t} = (m-1)^{-1} \sum_{j \in M} d_{ij,t} \qquad i = 1,\ldots,m, \tag{4.23}$$

represents the simple loss of model i relative to any other model j in the set at time t. For the EPA statistical tests, the hypothesis can be constructed in two alternative ways:

$$\begin{aligned} H_{O,M} &: c_{ij=0}, && \forall i,j = 1,2,\ldots,m \\ H_{A,M} &: c_{ij\neq 0}, && \exists i,j = 1,\ldots,m \end{aligned} \tag{4.24}$$

or

$$\begin{aligned} H_{O,M} &: c_{i=0}, && \forall i = 1,2,\ldots,m \\ H_{A,M} &: c_{i\neq 0}, && \exists i = 1,\ldots,m \end{aligned} \tag{4.25}$$

where $c_{ij} = \mathbb{E}(d_{ij})$ and $c_i = \mathbb{E}(d_i)$. Following Hansen, Lunde, and Nason (2011), to test the two hypothesis above, we construct two statistics:

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{var}(\bar{d}_{ij})}} \qquad \text{and} \qquad t_{i\cdot} = \frac{\bar{d}_{i,\cdot}}{\sqrt{\widehat{var}(\bar{d}_{i,\cdot})}} \qquad \text{for} \quad i,j \in M, \tag{4.26}$$

where $\bar{d}_{i,\cdot} = (m-1)^{-1}\sum_{j\in M}\bar{d}_{ij}$ represents the simple loss of the i-th model relative to the averaged losses across models in the set M, and $\bar{d}_{ij} = (m)^{-1}\sum_{t=1}^{m}\bar{d}_{ij,t}$ measures the relative sample loss between the i-th and j-th models (Bernardi and Catania 2018). The denominators, $\widehat{var}(\bar{d}_{i,\cdot})$ and $\widehat{var}(\bar{d}_{ij})$ are the bootstrapped estimates variances of $var(\bar{d}_{i,\cdot})$ and $var(\bar{d}_{ij})$, respectively. To calculate the bootstrapped variances, we follow Hansen, Lunde, and Nason (2011) and perform a block-bootstrap procedure of 5000 resamples. The block length p is calculated as the max number of significant parameters obtained after an AR(P) process is fit on all the d_{ij} terms. The first t-statistic, t_{ij}, is a well known statistic when comparing two forecasts; see Diebold and Mariano (2002) and West (1996). The second t-statistic, t_i, is used in Hansen, Lunde, and Nason (2003), P. R. Hansen and Lunde (2005), and Hansen, Lunde, and Nason (2011). The two EPA null hypotheses represented in equation (4.24) and (4.25) map naturally into two test statistics:

$$T_{R,M} = max_{i,j \in M}|t_{ij}| \qquad \text{and} \qquad T_{max,M} = max_{i \in M} t_{i\cdot}, \tag{4.27}$$

As per the discussion around the sequential procedure of the MCS process, the choice of the worst model to be eliminated, on grounds of the rejection of EPA, is made using an elimination rule that is coherent with the statistic test defined in equation (4.26):

$$e_{max,M} = \arg\max_{i \in M} \frac{\bar{d}_{i,\cdot}}{\widehat{var}(\bar{d}_{i,\cdot})}, \qquad e_{R,M} = \arg\max_{i} \left\{ \sup_{j \in M} \frac{\bar{d}_{ij}}{\sqrt{\widehat{var}(\bar{d}_{ij})}} \right\}, \tag{4.28}$$

respectively. We use a p-value of 0.05 representing a 95% certainty of elimination.

4.6 Results

In order to generate the forecasts, 100 000 draws are made from the predictive posterior distribution of the BVAR at each t for all forecast horizons $h = \{1, 2, 3, 4, 6, 8\}$, for each of the personal consumption expenditure components. Figure 4.7 and 4.8 visualise the forecast performance comparison among the models for different forecast evaluation measures (CRPS, LS and RMSE), separating short-term ($t + 1$ and $t + 2$) and medium-term ($t + 3$ and $t + 4$) forecasts:[14]

[14] The visual results for periods $t + 6$ and $t + 8$ can be requested from the author.

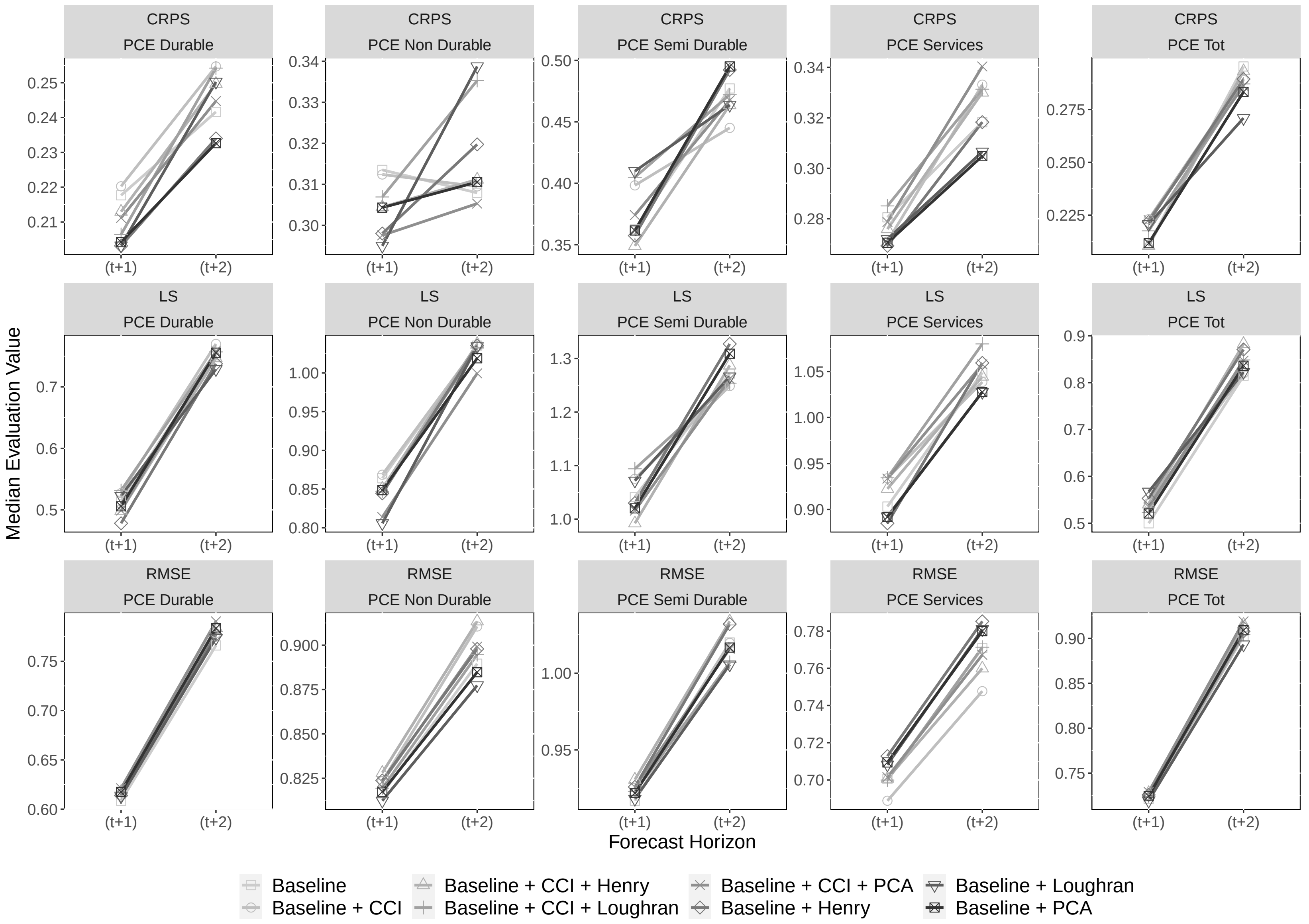


Figure 4.7: Comparison of evaluation metrics of different models on short-term forecasting horizons $t+1$ and $t+2$ using evaluation metrics CRPS, LS, and RMSE.

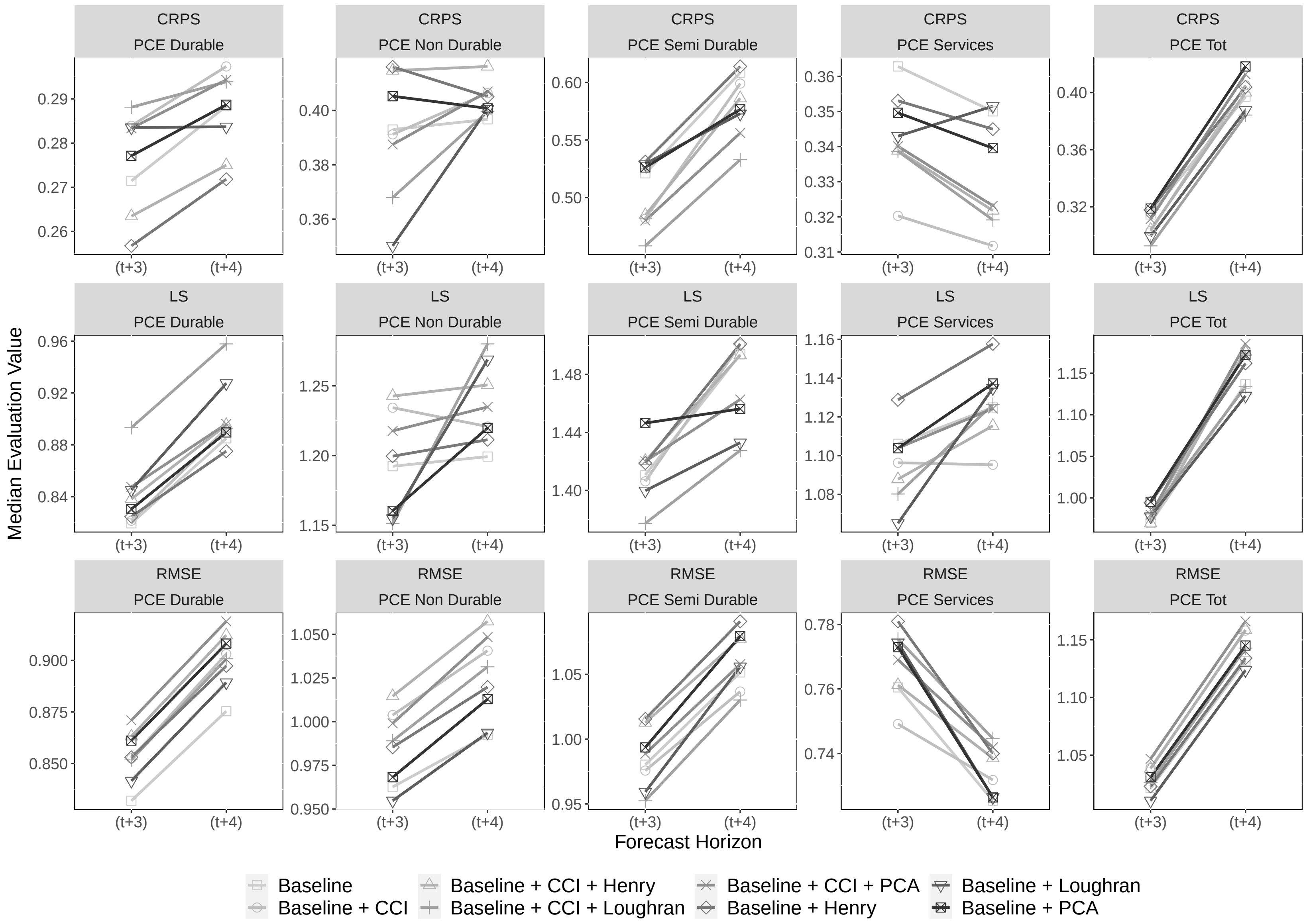
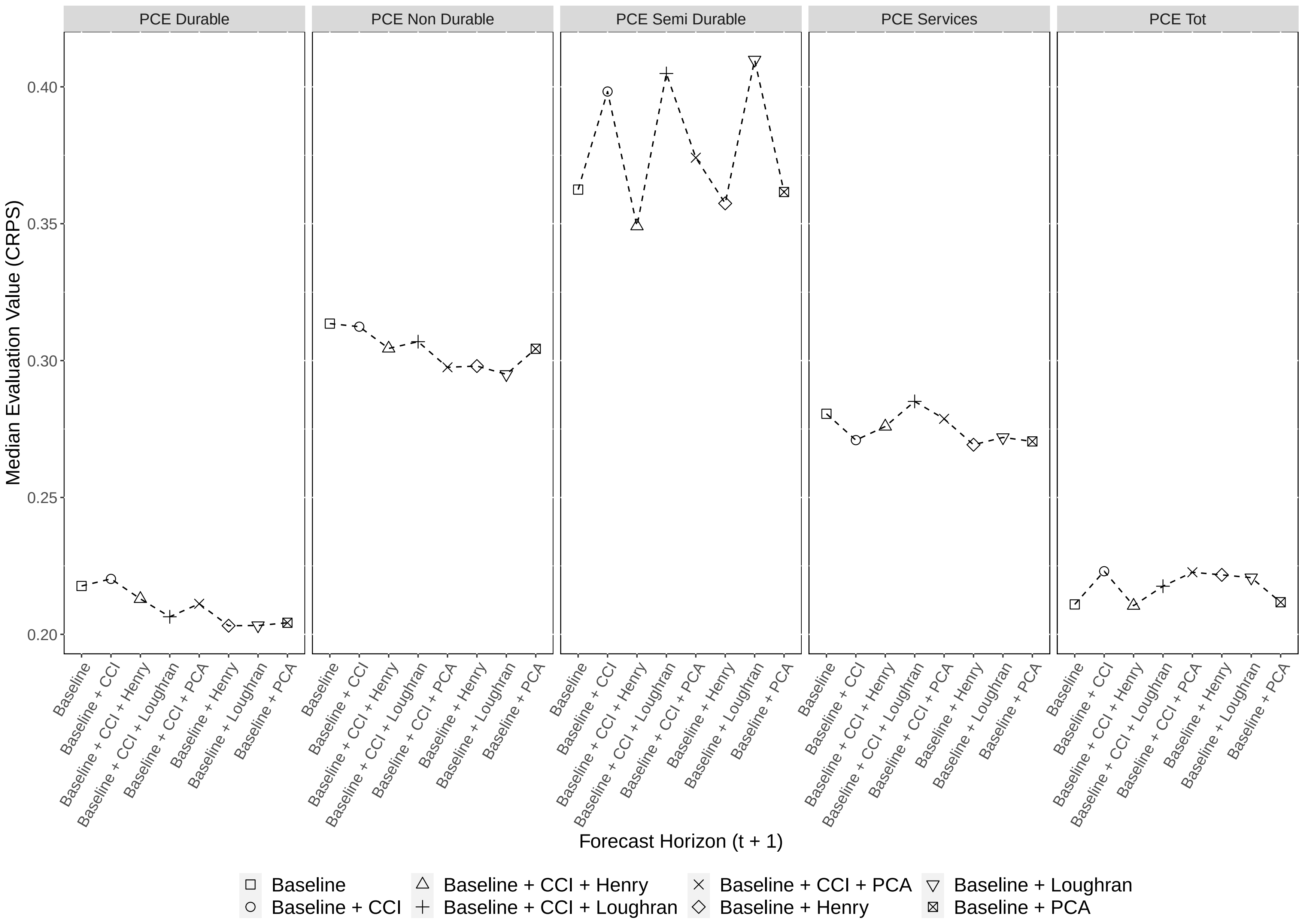
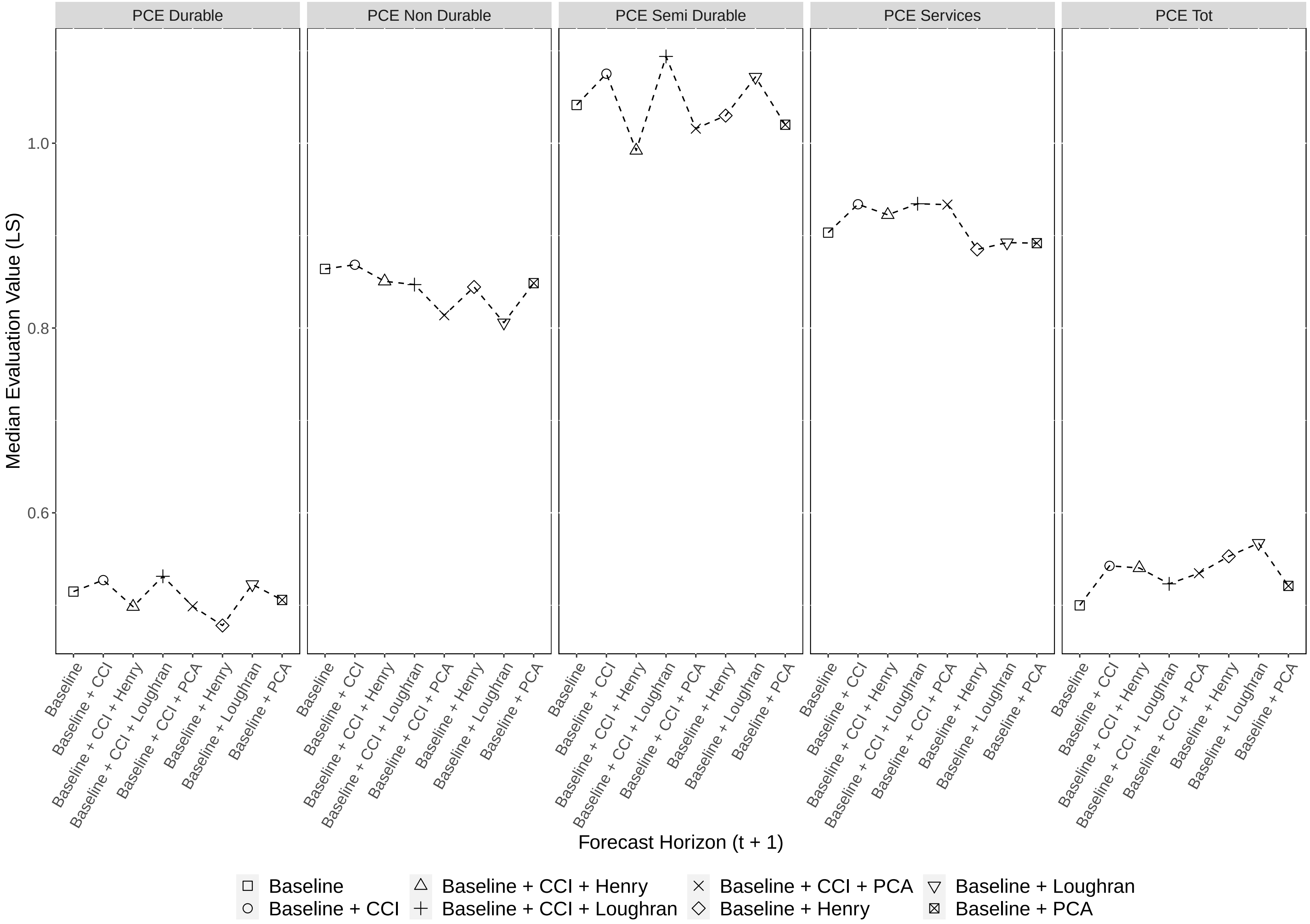


Figure 4.8: Comparison of evaluation metrics of different models on medium-term forecasting horizons $t+3$ and $t+4$ using evaluation metrics CRPS, LS, and RMSE.

Figure 4.7 displays a stark difference in the forecasting performance for the forecast horizon $h = 1$ and $h = 2$. The models can be seen to perform almost the same in $t+1$, only dispersing in forecast period $t+2$. This can be misleading as the rate of change in performance measures in the short term changes quite drastically when comparing $t+1$ with $t+2$ forecast metrics. For the LS and CRPS scores, the range of the forecast performance can vary as much as 30% between the periods. We also see a large range between periods for RMSE evaluation measures in the short-term forecasts. To highlight the key difference between the models at forecast horizon $h = 1$, figure 4.9 (a) - (c) plots the within-period forecast performance. The figure represent the median LS and CRPS scores (and RMSE) for the different models. As is the case with most loss functions, the lower the score, the better the forecasting performance of a model.


PCE Durable
PCE Non Durable
PCE Semi Durable
PCE Services
PCE Tot
Median Evaluation Value (CRPS)
0.20
0.25
0.30
0.35
0.40
Baseline
Baseline + CCI
Baseline + CCI + Henry
Baseline + CCI + Loughran
Baseline + CCI + PCA
Baseline + Henry
Baseline + Loughran
Baseline + PCA
Forecast Horizon (t + 1)
□ Baseline
○ Baseline + CCI
△ Baseline + CCI + Henry
+ Baseline + CCI + Loughran
× Baseline + CCI + PCA
◇ Baseline + Henry
▽ Baseline + Loughran
⊠ Baseline + PCA


PCE Durable
PCE Non Durable
PCE Semi Durable
PCE Services
PCE Tot
Median Evaluation Value (LS)
1.0
0.8
0.6
Baseline
Baseline + CCI
Baseline + CCI + Henry
Baseline + CCI + Loughran
Baseline + CCI + PCA
Baseline + Henry
Baseline + Loughran
Baseline + PCA
Forecast Horizon (t + 1)
□ Baseline
○ Baseline + CCI
△ Baseline + CCI + Henry
+ Baseline + CCI + Loughran
× Baseline + CCI + PCA
◇ Baseline + Henry
▽ Baseline + Loughran
⊠ Baseline + PCA

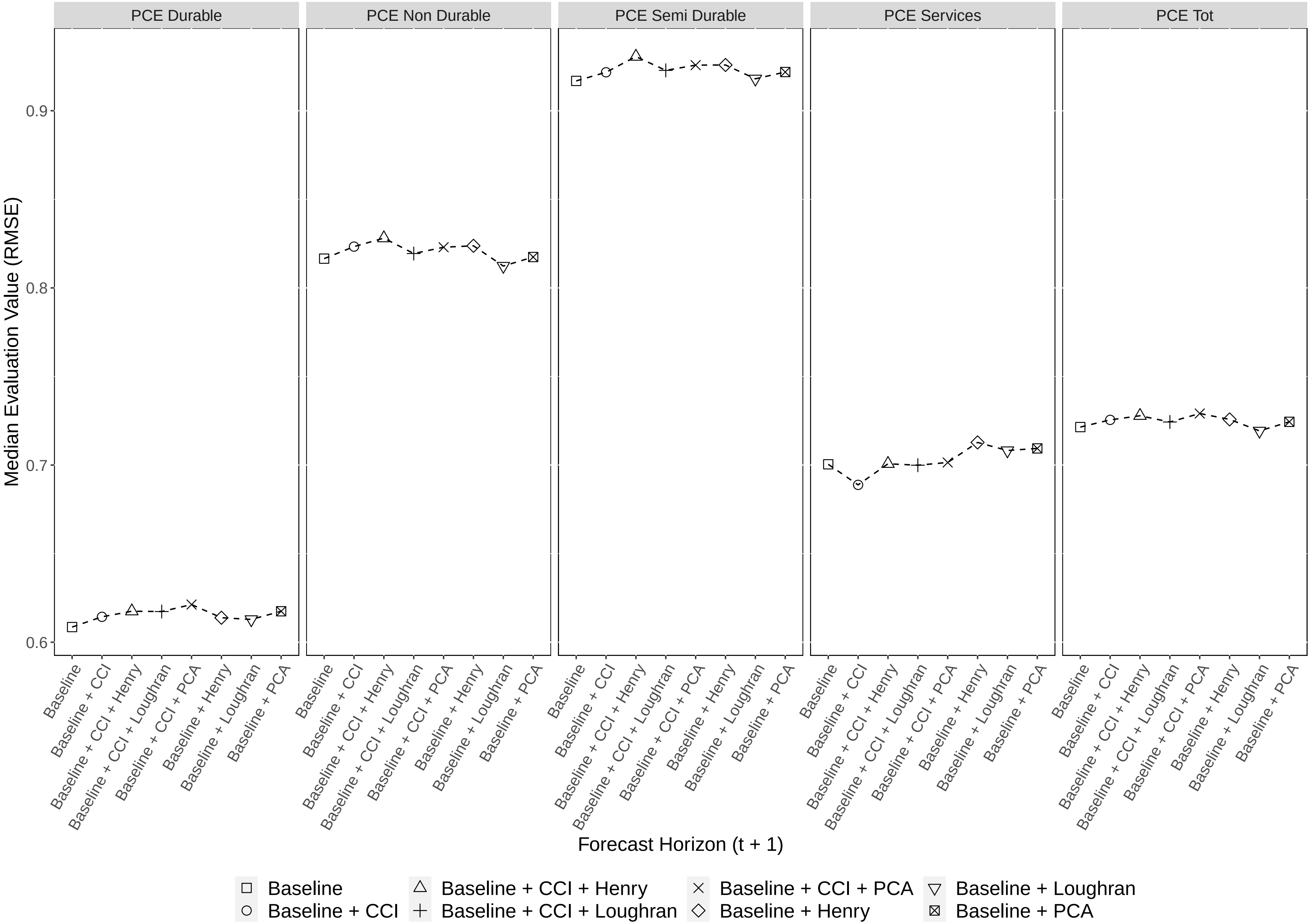


Figure 4.9: Comparison of evaluation metrics of different models with forecast horizon $h = 1$. The figure illustrates within-period performance of the models.

Viewing the within-period performance for the short term, $t+1$, forecast, a clearer idea on model performance is obtained. For the total personal consumption expenditure series, the combination model Baseline + CCI + Henry, the PCA media-based index, and the baseline provide better pseudo out-of-sample forecasts than CCI on its own. In fact, the CCI does the worst in terms of forecasting performance for total PCE when only taking the CRPS score into account. We also observe that the MSIs perform better in terms of CRSP and LS scoring for three out of the five PCE measures: Durables, Non-durables, and Services. The point estimate RMSE measure is in agreement with the density-based measures in that the media-based sentiment indices perform well in the short horizon forecast. The RMSE measure especially highlights the Baseline + Loughran model's forecasting performance in forecasting Non-durables, Semi-durables and total PCE.

Besides the $t+1$ forecast, the chapter also investigates the forecast performance for $t+2$. In terms of CRPS for total PCE, the Loughran-based indices outperform all other model specifications. The media-based indices also perform well in forecasting Durables and Services in $t+2$, which is in agreement with the $t+1$ results.

Table B.5 gives the forecast ranking (1 - best, 8 - worst) of each model per PCE for the periods $h = \{1 \ldots 8\}$, while table 4.2 displays the median CRPS values at the different horizons. For the rest of the chapter, the CRPS score will be considered the primary measure of predictive performance due to its attractive properties.

Table 4.2: Median CRPS values as a measure of forecast model performance at different horizons listed in section 4.5.1.

Outcome	Horizon	Baseline	Baseline + CCI	Baseline + CCI + Henry	Baseline + CCI + Loughran	Baseline + CCI + PCA	Baseline + Henry	Baseline + Loughran	Baseline + PCA
PCE Durable	(t+1)	0.218	0.220	0.213	0.206	0.211	**0.203**	0.203	0.204
PCE Durable	(t+2)	0.242	0.255	0.250	0.254	0.245	0.234	0.250	**0.233**
PCE Durable	(t+3)	0.271	0.284	0.263	0.288	0.283	**0.257**	0.284	0.277
PCE Durable	(t+4)	0.288	0.297	0.275	0.294	0.294	**0.272**	0.284	0.289
PCE Durable	(t+6)	0.318	0.334	0.327	0.373	0.342	**0.315**	0.340	0.322
PCE Durable	(t+8)	**0.306**	0.322	0.338	0.349	0.330	0.326	0.335	0.319
PCE Non Durable	(t+1)	0.314	0.312	0.304	0.307	0.298	0.298	**0.295**	0.304
PCE Non Durable	(t+2)	0.308	0.309	0.311	0.335	**0.305**	0.320	0.339	0.311
PCE Non Durable	(t+3)	0.393	0.391	0.415	0.368	0.387	0.416	**0.350**	0.405
PCE Non Durable	(t+4)	**0.397**	0.406	0.416	0.400	0.407	0.405	0.401	0.401
PCE Non Durable	(t+6)	0.436	0.443	0.460	0.422	0.439	0.434	**0.403**	0.431
PCE Non Durable	(t+8)	0.465	0.484	0.488	0.441	0.479	0.469	**0.434**	0.453
PCE Semi Durable	(t+1)	0.362	0.398	**0.349**	0.405	0.374	0.357	0.410	0.362
PCE Semi Durable	(t+2)	0.477	**0.445**	0.464	0.472	0.468	0.492	0.464	0.495
PCE Semi Durable	(t+3)	0.521	0.482	0.485	**0.458**	0.480	0.531	0.529	0.526
PCE Semi Durable	(t+4)	0.608	0.599	0.586	**0.533**	0.556	0.614	0.573	0.577
PCE Semi Durable	(t+6)	0.504	0.501	0.527	**0.485**	0.511	0.512	0.492	0.512
PCE Semi Durable	(t+8)	**0.505**	0.528	0.571	0.552	0.564	0.537	0.515	0.519
PCE Services	(t+1)	0.281	0.271	0.276	0.285	0.279	**0.269**	0.272	0.270
PCE Services	(t+2)	0.318	0.333	0.330	0.331	0.340	0.318	0.307	**0.305**
PCE Services	(t+3)	0.363	**0.320**	0.339	0.339	0.340	0.353	0.343	0.350
PCE Services	(t+4)	0.350	**0.312**	0.322	0.319	0.323	0.345	0.352	0.340
PCE Services	(t+6)	0.339	0.344	0.348	**0.336**	0.350	0.355	0.354	0.350
PCE Services	(t+8)	**0.335**	0.351	0.360	0.365	0.359	0.351	0.368	0.357
PCE Tot	(t+1)	0.211	0.223	**0.210**	0.218	0.223	0.222	0.221	0.212
PCE Tot	(t+2)	0.295	0.289	0.293	0.287	0.286	0.289	**0.271**	0.283
PCE Tot	(t+3)	0.315	0.303	0.304	**0.293**	0.311	0.318	0.299	0.319
PCE Tot	(t+4)	0.397	0.405	0.400	**0.384**	0.413	0.404	0.388	0.418
PCE Tot	(t+6)	0.418	0.426	0.436	0.413	0.455	0.426	**0.406**	0.441
PCE Tot	(t+8)	**0.399**	0.401	0.409	0.420	0.417	0.403	0.413	0.407

The results in table B.5 and 4.2 suggest that consumer confidence measures (albeit text- or survey-based), as an explanatory variable, in the modelling of personal expenditure is effective in both the short and the long run. The improvement in the forecasting of PCE and its subcomponents above that of the baseline model agrees with the results of Wilcox (2007) who also found that by incorporating a consumer confidence measure, forecasting accuracy improved. This can be seen in figure 4.9 and table B.5 where the best-ranked models included either the traditional or media-based measures. The results also highlight the advantage of the media-based sentiment indices. The media indices outperform their traditional counterpart for several disaggregations of PCE when investigated at $h = 1$ and $h = 2$ forecast horizons. The finding indicates that forecasting of personal consumption expenditure could benefit from the inclusion of a news-driven index in its modelling and forecasting framework. We proceed to statistically construct superior model sets using the MCS procedure employing CRSP scores as the loss function to determine the best model (or set of models) for each of the personal expenditure measures at the different horizons.

Table 4.3: Proportional composition of models remaining in the SSM after applying the MCS procedure.

Model	N	Prop Count	Cumulative Sum
Baseline + Loughran	17	30.9%	30.9%
Baseline	13	23.6%	54.5%
Baseline + CCI	8	14.5%	69.1%
Baseline + CCI + Loughran	6	10.9%	80%
Baseline + PCA	6	10.9%	90.9%
Baseline + CCI + PCA	3	5.5%	96.4%
Baseline + CCI + Henry	1	1.8%	98.2%
Baseline + Henry	1	1.8%	100%

The MCS procedure highlights the effectiveness of the media-based sentiment indices in the forecasting of personal consumption expenditure measures. Of the 55 models that the MCS procedure isolated, 60% or 34 of the models were exclusively media-based or a combination of media-based and traditional CCI models.

Given the advantage of the MSI in modelling PCE in the short term, in the last part of the analysis, the chapter investigates whether a media-based index can outperform its rival CCI index in times of crisis. The hypothesis is that in times of crisis, the news media will be quicker to update and consequently should be able to capture market confidence more efficiently. To investigate this hypothesis, we looked into how the CCI and MSI performed during the financial crisis. The crisis will be defined as July 2007 to July 2009. For this period, we compared the actual and the estimate of the models for each of the different PCE measures for all forecasts $t + 1$. In times of financial uncertainty, it is imperative that decision-makers have access to forecasts that reflect the true nature of the underlying economy. For the financial crisis period, the best model is considered to be the model with the lowest CRPS score.

Table 4.4: Density forecast evaluation measures as per CRPS for the period of the Financial crisis (July 2007 - July 2009).

Outcome	Models	2007-09-01	2007-12-01	2008-03-01	2008-06-01	2008-09-01	2008-12-01	2009-03-01	2009-06-01
PCE Durable	Baseline	0.167	0.977	0.611	0.321	0.427	0.143	0.130	0.967
PCE Durable	Baseline + CCI	0.195	1.014	0.652	0.290	0.392	0.142	0.130	**0.936**
PCE Durable	Baseline + CCI + Henry	0.191	1.037	0.673	0.297	0.412	**0.140**	0.129	0.993
PCE Durable	Baseline + CCI + Loughran	0.184	0.989	0.636	**0.262**	**0.377**	0.145	0.129	1.013
PCE Durable	Baseline + CCI + MSI	0.188	1.016	0.660	0.287	0.402	0.143	**0.128**	1.022
PCE Durable	Baseline + Henry	0.177	1.031	0.661	0.330	0.436	**0.140**	0.129	1.012
PCE Durable	Baseline + Loughran	**0.164**	**0.972**	**0.607**	0.294	0.411	0.147	**0.128**	1.039
PCE Durable	Baseline + MSI	0.172	1.010	0.642	0.323	0.432	0.145	0.129	1.046
PCE Non Durable	Baseline	0.154	1.901	1.837	0.505	0.395	0.355	0.403	0.217
PCE Non Durable	Baseline + CCI	0.158	1.958	1.916	0.473	0.378	0.363	**0.390**	0.219
PCE Non Durable	Baseline + CCI + Henry	0.156	1.981	1.937	0.489	0.411	0.354	0.391	0.222
PCE Non Durable	Baseline + CCI + Loughran	0.155	1.915	1.860	**0.395**	**0.369**	0.326	0.437	**0.198**
PCE Non Durable	Baseline + CCI + MSI	0.155	1.964	1.896	0.462	0.407	0.342	0.411	0.206
PCE Non Durable	Baseline + Henry	**0.151**	1.958	1.902	0.527	0.411	0.357	0.400	0.230
PCE Non Durable	Baseline + Loughran	0.152	**1.883**	**1.796**	0.449	0.388	**0.315**	0.447	0.199
PCE Non Durable	Baseline + MSI	**0.151**	1.939	1.859	0.525	0.420	0.338	0.425	0.209
PCE Semi Durable	Baseline	**0.208**	0.235	0.345	0.354	0.262	0.700	0.704	0.241
PCE Semi Durable	Baseline + CCI	0.216	0.231	0.319	0.331	0.250	**0.681**	0.738	0.248
PCE Semi Durable	Baseline + CCI + Henry	0.215	0.234	0.327	0.349	0.275	0.709	0.738	0.230
PCE Semi Durable	Baseline + CCI + Loughran	0.211	**0.221**	**0.304**	**0.275**	**0.234**	0.699	0.666	0.210
PCE Semi Durable	Baseline + CCI + MSI	0.214	0.231	0.317	0.321	0.258	0.697	0.698	0.214
PCE Semi Durable	Baseline + Henry	0.216	0.256	0.379	0.376	0.276	0.715	0.700	0.229
PCE Semi Durable	Baseline + Loughran	0.209	0.241	0.356	0.307	0.243	0.707	**0.618**	**0.206**
PCE Semi Durable	Baseline + MSI	0.214	0.255	0.376	0.359	0.268	0.706	0.651	0.214
PCE Services	Baseline	0.230	0.237	0.377	0.367	**0.224**	1.675	1.036	**0.239**
PCE Services	Baseline + CCI	0.249	0.227	0.341	0.336	0.229	**1.592**	**0.927**	0.240
PCE Services	Baseline + CCI + Henry	0.249	0.227	0.349	0.324	0.227	1.618	0.978	0.249
PCE Services	Baseline + CCI + Loughran	0.255	0.227	**0.325**	**0.292**	0.244	1.611	1.006	0.244
PCE Services	Baseline + CCI + MSI	0.252	**0.226**	0.337	0.311	0.233	1.615	0.988	0.245
PCE Services	Baseline + Henry	**0.229**	0.237	0.386	0.348	0.227	1.695	1.043	0.251
PCE Services	Baseline + Loughran	**0.229**	0.238	0.372	0.329	0.240	1.698	1.099	0.248
PCE Services	Baseline + MSI	**0.229**	0.240	0.381	0.349	0.230	1.688	1.068	0.247
PCE Tot	Baseline	**0.133**	**1.108**	**0.692**	**0.130**	0.424	0.988	**0.480**	0.649
PCE Tot	Baseline + CCI	0.161	1.155	0.731	0.135	**0.407**	0.952	0.486	0.641
PCE Tot	Baseline + CCI + Henry	0.160	1.152	0.734	0.135	0.410	0.961	0.507	0.633
PCE Tot	Baseline + CCI + Loughran	0.156	1.150	0.725	0.135	0.413	**0.944**	0.501	0.636
PCE Tot	Baseline + CCI + MSI	0.158	1.153	0.734	0.133	0.426	0.954	0.507	0.632
PCE Tot	Baseline + Henry	0.138	1.125	0.709	0.132	0.410	0.992	0.503	**0.631**
PCE Tot	Baseline + Loughran	**0.133**	1.113	0.701	**0.130**	0.425	0.983	0.491	0.640
PCE Tot	Baseline + MSI	0.136	1.122	0.710	**0.130**	0.433	0.991	0.500	0.635

Table 4.4 shows the forecasting performance of all the models during the financial crisis. Overall, the Loughran-based indices did well during this period, achieving the best forecasts for Durables, Non-durables, and Semi-durables. The results for Services agree with the MCS results; models that include the CCI as a predictor variable performed the best during the financial crisis, while the media-based indices performed the worst. For total personal consumption expenditure during the financial crisis, the baseline came out as the top predictive model.

The findings could be an indication of how the news component was driving expenditure during this time, while the confidence indicator had to catch up to how fast the economy was deteriorating. Midst the financial crisis, disposable income and expenditure tightened at a quicker pace than the quarterly sentiment indicators could capture severely affecting durable spend. This aligns with the findings of Beaudry and Portier (2004) and Beaudry and Portier (2014) where changes in agents' information, due to the arrival of news, can cause business cycle fluctuations driven by expectational change. Similarly, in the financial crisis, as the news started playing a more central role in being an indicator of the economy it exacerbated the downturn and households quickly tightened their expenditure behaviour.

4.7 Conclusion

The results from this chapter support findings from literature, that make use of survey-based consumer and business confidence indicators, in that confidence indicators (whether text- or survey-based) contain information for forecasting real economic activity which is not contained in other economic variables (Santero and Westerlund 1996; Ludvigson 2004; Kabundi, Nel, and Ruch 2016). The evidence provided suggests that not only could media-based consumer confidence indices help to improve the forecasting of key economic indicators such as personal consumption expenditure, but it could also improve the performance of forecasting models at short and longer horizons. This finding corroborates results from Wilcox (2007), which found that consumer confidence improved the forecasting error for a period of up to four quarters.

We illustrated this by estimating different constructions of a Bayesian VAR forecasting model which include various combinations of sentiment dictionaries along with a baseline specification that conditions for macroeconomic fundamentals. These specifications were used to forecast personal consumption expenditure as well as its sub-components for forecast horizons $h = \{1, 2, 3, 4, 6, 8\}$. We evaluated the out-of-sample forecasting performance of each of the specifications through density-based scoring models: CRPS and LS (and a point estimate RMSE). The chapter also evaluated the forecasting performance statistically through a procedure called Model Confidence Sets. Using the median CRPS score as the primary performance measure, the results showed that the inclusion of a Loughran-based sentiment index into the forecasting model improved forecasts for Durables, Non-durables, and total PCE.

The statistical MCS procedure showed that the Loughran-based model specification could greatly help improve the forecasting of total PCE, but improving forecasting performance is less promising for the sub-components. For Durables, the baseline model does well across horizons, while for Semi-durables, the media-based indices could help improve forecasts at a longer horizon. The MCS procedure also revealed how the median CRPS score could be misleading in determining forecasting performance. Using the median CRPS score, the Loughran-based dictionary can be considered the best model for multiple sub-components of PCE, while the statistical procedure that incorporates block bootstrapping shows how this only holds for total PCE. Lastly, the chapter investigated the performance of the different sentiment indices during the financial crisis. Here, the Loughran dictionary-based models did well to improve the $t+1$ forecasts of the sub-components of PCE within the period, but struggled to improve on the baseline for the overall changes in total PCE. The empirical results in this thesis motivate for further research into how text-based analysis could contribute to improving (or creating new) official economic indicators.

Chapter 5

Domain specific dictionary generation using machine learning

...but the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem. - Leo Breiman

This chapter aims to offer an alternative to the manual, labour-intensive process of constructing a domain-specific lexicon or dictionary through the operationalisation of subjective information processing. Traditionally, in order to analyse the tone of any news texts, computational linguistics employs manually pre-selected dictionaries. These dictionaries range from general (Harvard-IV, Liu (2012) and Nielsen (2011)) to more domain-specific lexicons such as those of Loughran and McDonald (2011) and Henry (2008) which have a financial focus. However, there remains a substantial difference in the word lists of these domain-specific dictionaries. Despite the range of dictionaries to choose from, one dictionary alone might not be sufficient to capture the nuances found within a specific news domain. The choice of a suitable dictionary poses a challenge in itself, even before any analysis has started.

Machine learning is presented as one way to create such a domain-specific dictionary. This approach would enable the creation of dictionaries tailored to a specific need in an automated fashion. Being less subjective, these dictionaries are also more easily tested and replicable. Tokens (words) are statistically selected using recursive feature elimination to generate a domain-specific dictionary from a corpus of text. An individual domain-specific sentiment lexicon and respective sentiment index is generated for the CCI, BCI, Building Confidence index, Civil Confidence index, PMI, as well as the leading indicator constructed by the South African Reserve bank (Leading). To accomplish this, we applied a random forest algorithm to select the most important words from a corpus of business news media, determined whether the token has a positive or negative relationship with the outcome through ordinary least squares, and constructed a corresponding index.

The Random forests (RF) technique is increasingly used in a range of fields due to its high prediction ability and the feature selection procedure it inherently contains (Biau and D'Elia 2009; Meinshausen 2006; Athey, Tibshirani, and Wager 2016). Although RFs have been widely applied in bioinformatics and related fields, very few applications within economics have seen the use of the algorithm in time-series problems. This is mainly due to the fact that the algorithm assumes each data point to be independent and so violates the time dependency assumption within time-series data. Despite the algorithm's flaws in terms of pure prediction or forecasting in time-series problems, RFs can still be used as a feature selection machine where the curse of dimensionality, $p >> n$, is a problem (Kane et al. 2014; Tyralis and Papacharalampous 2017).

Building on current empirical literature that has explored the media-economic nexus, we contribute by (a) constructing a domain-specific dictionary for various economic confidence indices, (b) introduce a novel weighting schema of text tokens that aim to account for time dependence, and (c) operationalise subjective information processing of text data using machine learning.

The machine-generated dictionaries contain both unigrams and bigrams as features, creating a richer feature space in comparison to unigram tokens. Sentiment indices constructed from machine-generated dictionaries are shown to have a better fit with the multiple indicators investigated when compared to the sentiment index constructed from a commonly used financial dictionary. The domain-specific sentiment indices also show a significantly lower root mean squared error (RMSE) in the five-year holdout sample period from 2012 to 2017. These results support the case for domain-specific dictionaries being able to pick up nuances found within domain-specific topic news. The results also suggest that having a manually generated dictionary act as a prior narrows the tokens that the Random

Forest has to search over, while maintaining the same lower RMSE out-of-sample as an unrestricted model that includes all the tokens. Employing a manually generated dictionary such as that of Loughran and McDonald (2011) also decreases the computation burden on the pre-processing[1] and estimation of the dictionaries.

This chapter starts off by examining the relationship between expectations and the role it plays in shaping how the economy evolves. Section 5.1 highlights the reason why expectations matter by illustrating that expectations are able to provide foresight and as such play an important part in policy design. The section draws from the inflation expectation literature to provide a theoretical basis for how the media forms expectations which, in turn, influence the real economy. We also expand how computational linguistics is applied within economics to derive economic sentiment, especially utilising text from media sources. The section ends by providing an overview of the current methods and techniques that are utilised to create domain-specific dictionaries. Following an extensive overview of the theoretical arguments and relevant literature, section 5.2 discusses the framework to create domain-specific dictionaries. This section focuses on the random forest algorithm that is used, as well as the subjective choices that have to be made in order to generate a domain-specific dictionary using the proposed methods. Section 5.3 follows to provide a thorough evaluation of the domain-specific dictionaries and the respective sentiment indices created from them. The dictionaries are compared in terms of the tokens each of the dictionaries consists of. We also showcase the specific bigrams that were selected by the algorithm as a network analysis. After the dictionaries' constructs are discussed, each of their resulting sentiment indices is evaluated against various confidence indices using RMSE as a measure to determine the best fit. Lastly, section 5.4 ends with the conclusion of the chapter and suggestions for future research.

[1] Pre-processing is that step in which the data gets transformed, or encoded, to bring it to such a state that the machine can easily parse the features of the data.

5.1 The news, the economy, and text analysis

5.1.1 The news and economy

The study of the closely related relationship between economic reporting and the real economy is not new in terms of scholarly attention. Early work such as that of Goidel and Langley (1995) already identified a link between news coverage and how it influences both the perception and evaluation of the real economy. Goidel and Langley (1995) were also some of the first researchers to posit that the media tends to follow negative economic conditions more closely and is accordingly characterised by a set of persistent biases. Given this known bias, it is important to also acknowledge that the media exercises plenty of latitude in what they deem to be important. This entails that the relationship between the economy and economic media is not stable, but dynamic. This dynamic relationship between the economic agent and news media can also be seen in the work of Carroll (2003).

The authors illustrate that although empirical household expectations cannot be deemed to be rational, expectation dynamics can be explained through a model that incorporates professional forecasters' views. In this model, the household's views and expectations are influenced through news reports which reflect on the views of professional forecasters, who, in turn, could be considered to be rational agents. The agents absorb macroeconomic trends and economic content by using a model of probabilistic absorption of news stories as opposed to assimilating raw official statistical. This inattention to news stories by choice creates a stickyness in aggregate expectations that, in turn, have important macroeconomic consequences. Carroll (2003)'s baseline model follows a very similar approach to Mankiw and Reis (2002) in modelling the evolution of mean expectations. Apart from finding that expectations are sticky due to rational inattention (see Sims (2003)), the authors also find that the data showed a preference for a forward-looking version of the model. This is in contradiction with an adaptive expectations model whereby expectations are adjusted in line with recently reported statistics. The results of the paper are important as they act as a theoretical model in explaining and motivating that news reports impact forward-looking expectations. This model provides a plausible middle ground between fully rational expectations and adaptive expectations.

The stickyness of expectations is further explored through a labour market model proposed by Akerlof et al. (2000). Here, the economic agents only concern themselves with inflation when the ignorance thereof will become costly. This rational inattention model can also be applied in the context of news sentiment as discussed further below. In the inflation framework, when inflation is low, it is not very salient and the relationship between expected inflation and the wage and price setting is not as strong. But as inflation rises, so does the importance of anticipating it correctly, resulting in the price and wage setting fully responding to inflation expectations. Given that the largest expense for most businesses would be wages, if the household's expectations of inflation translate to a change in nominal wage demand, the firm's pricing decisions will be affected through the usual wage-push channel. Although the information contained in news reports has an important role to play in adjusting expectations such as inflation, the volume and bias of the media can in itself impact the real economy. Lamla and Lein (2014) examines the role of news media and how it influences consumer expectations around inflation. They emphasise the role of information rigidities and how the media plays an important part in transmitting information about the macroeconomic conditions to the consumer. Through the use of a theoretical model, the authors show that the media affects consumers and expectations mainly through two channels: volume and bias. The higher the volume of news available to the agents in the economy, the more accurate their expectations become; but this effect could be reversed if the news is consistently biased. Using a dataset of German news articles, the results show that given a higher volume of information, agents have a higher propensity to update their expectation, resulting in more accurate forecasts. The empirical results also provide evidence that a one standard deviation increase in neutrally toned media improved consumer's expectations by around 20%, while the increase of negatively toned media deteriorated the results by roughly the same amount. This direct effect on consumer's expectations also has an indirect effect on the real economy. Inflation expectations have a tendency to be self-fulfilling (Leduc, Sill, and Stark 2007). As inflation expectations keep adjusting upwards, the effectiveness of monetary policy could be impeded. This interplay between the flow of information through the news media and the real economy is not only observed in the adjustment of inflation expectations, but also in the adjustment of borader macroeconomic expectations.

Nadeau et al. (1999) evaluated the process by which business elites' expectations and retrospectations trickle through to news media, eventually impacting not only the economy but also matters such as presidential approval. By using content analysis, the results showed that although information is transmitted through news media channels, it is not always completely unchanged. The media is partially autonomous, acting as a mediator between expert opinion and the mass public on economic conditions. The author's research builds on the idea of media-dependency theory in which the experts' economic views will be widely and correctly reported by journalists. For the mass public, this reported information is by far the best information on the future state of the economy. But, with economic reporting overemphasising poor economic news, the public does not always adjust their expectations contemporaneously due to rational inattention, as it is a costly exercise to do so.

Although research such as that of Carroll (2003), Akerlof et al. (2000), and Lamla and Lein (2014) focuses on how the media can influence expectations around the inflation that eventually flows through to the real economy, Nadeau et al. (1999) demonstrate how general economic coverage by the news media can subsequently affect public opinion about any economic or political outcome in a similar manner. This in part questions the extent of the influence of news media and if, in a rational inattention model, the news says something about the past, present, or future of the economy. Soroka, Stecula, and Wlezien (2015) explored this question empirically.

They asked the question of whether media news coverage has a forward-looking or retrospective agenda towards economic sentiment. To test this, Soroka, Stecula, and Wlezien (2015) relied on three sources of data: macroeconomic measures, media data, and measures of public opinion. The data spans the period 1980 to 2011. The macroeconomic data contains the leading, coincidental, and lagging indicator of the economy, which allows for the assessment of the impact of news media on each of the indicators. The indicators are purged from all sentiment measures so as not to introduce measure-induced endogeneity. The media data contains articles from the New York Times and Washington Post relating to major economic issues that had been transformed into indices indicating tone (using sentiment analysis) and volume of reporting. The relationship between the media and economy is tested using error correction models as they have the advantage of breaking down the long-term and short-term dynamics between series. Results from the model showed that the tone and volume of media reporting were significantly related to present and future economic activity, while the relationship with lagging indicators was insignificant. When the indicators were estimated in a saturated model (all indicators included), the leading indicator was the only significant variable. This indicates that conditioning on other indicators in the economy (ceteris paribus), the media was more concerned with reporting on the future. The authors also went on to evaluate whether media coverage had short-run or long-run impacts. Interestingly enough, the volume of reporting had long-term impacts, while the tone of reporting was only significant for the short-run coefficient. This would suggest that media content best reflects changes in the future economy rather than the level of economic conditions.

Contemporary literature extends the discussion around the media-economic connection by analysing text through what is known as text analysis or computational linguistics. In recent years, computational power has become more accessible, but at the same time the amount of information has increased drastically. This upsurge in information content can mainly be accredited to two large drivers: social media and online news media. Both of these factors have led to new avenues of research being developed to gain a better understanding on the role that the media plays in shaping expectations and the real economy. A large portion of text analysis within economics focuses on quantifying the sentiment of economic agents. This could be anything from deriving consumer and business sentiment through analysing news and blogs to investigating official central bank releases for information on macroeconomic policy direction.

Given the growing application of sentiment analysis to answer various economic questions, it is imperative that the right sentiment dictionary be used to analyse the text. Using machine learning algorithms to build domain-specific dictionaries could provide a possible avenue to create these dictionaries more efficiently in an automated fashion.

5.1.2 Machine learning to build domain-specific dictionaries

One of the earliest works on automatic dictionary generation involves polarising tokens, assuming words act the same as the spin of an electron. Takamura, Inui, and Okumura (2005)'s model extracts semantic orientation by comparing the directional spins of an electron (up or down) to the semantic orientation of words (positive or negative). Using the Spin Model and applying the mean field approximation, the authors compute the average orientation of each word. Using this method approximates the probability function of the system instead of computing the intractable actual probability function. This allows for the incorporation of more noisy data in the analysis. This was not previously possible with methods that focused on bootstrapping and shortest-path techniques to calculate semantic orientation. The proposed Spin Model outperformed both the bootstrap and shortest path in out-of-sample tests in terms of percentage precision. In more recent literature, text mining has started to apply machine learning methods to automatically construct a dictionary for sentiment analysis.

To better understand how companies frame their press releases, Prollochs, Feuerriegel, and Neumann (2015) constructed domain-specific dictionaries using announcement data. To do this, the authors used three different Bayesian approaches, namely LASSO, Ridge, and elastic net regression. These methods have been shown to have high explanatory power while also allowing for inference. All three of the models rely on Bayesian regularisation to conduct variable selection that shrinks the coefficients of non-informative variables. In the case of generating a dictionary, the model will converge to a parsimonious selection of tokens with high explanatory power. This approach permits for weights to be calculated on how strongly investors are influenced by information in the form of selected words. The authors do highlight that tokens identified as positive (negative) by the model, may not necessarily be interpreted positively by investors. The information contained in words is highly dependent on the context in which they appear. This means that one cannot assume that the model or manually generated polarity of a word necessarily equates to the linguistic orientation of the word (Loughran and McDonald 2011). The generated dictionaries are compared to existing financial dictionaries by evaluating their predictive performance for sentiment analysis on a validation set. Unlike previous methods, the announcements are not manually labelled; Prollochs, Feuerriegel, and Neumann (2015) instead use the corresponding stock market returns as the objective measure. The results showed that the ridge regression had the highest predictive performance out-of-sample, with an improvement of 93.25% increase in correlation in comparison to other well-known financial dictionaries.[2]

Machine learning has also been applied to forecasting key economic variables using newspapers. Using the text from three different newspapers, Kalamara et al. (2020) use a novel method of text counts, similar to the one used in this chapter, to extract timely signals to forecast GDP, CPI and unemployment. The findings showed that by using a supervised method that identifies relevant tokens, the method employed improved the forecasts when compared to existing text-based methods. Furthermore, the improvements were most prominent in periods of stress, where the importance of accurate forecasts are vital.

In contrast with the machine learning approach found in Prollochs, Feuerriegel, and Neumann (2015) and Kalamara et al. (2020), Labille, Gauch, and Alfarhood (2017) used probabilistic and information theory techniques to construct a domain-specific dictionary. By moving away from transferred supervised machine learning techniques, the method has the advantage of not having to update or adapt the constructed dictionary. To generate the three dictionaries, the authors used Amazon product reviews for 15 different categories submitted from January 2013 to July 2014. The reviews were star-rated 1 to 5. For their experiment, 1- and 2-star ratings were deemed negative, while 4- to 5-star reviews were labelled positive. This probabilistic approach was shown to outperform generic dictionaries with higher accuracy and F1-scores, indicating that for their experiment, domain-specific dictionaries were more accurate in the task of sentiment analysis.

[2] The correlation coefficient, ρ, was 0.1030 for the Ridge regression, while the Harvard-IV dictionary performed the best among the manually generated dictionaries with $\rho = 0.0533$.

5.2 Creating a domain-specific dictionary using random forests

This section provides an overview of the research methodology this authors implemented to generate domain-specific dictionaries using random forests.

Building on the findings of Odendaal, Reid, and Kirsten (2018), we analysed two financially oriented newspapers (Business Day and Financial Mail) to build a domain-specific dictionary.[3] The section is broken up in three subsections: how the token design matrix was constructed, explanation of the recursive feature elimination technique used to build the dictionaries and then an overview of practical considerations when using Random Forests is discussed.

5.2.1 Token weights

Before applying the random forest model, pre-processing steps were applied to the text. As is custom in any text analysis, the first step was to tokenise the text so that the columns present a token for a given t in time. Next, common words, often called stop words (including conjunctions), were removed from the text as these words contain little to no information value.

Once this procedure had been the completed, a suggested time-series weighting schema was applied to the design matrix of tokens. Following the premise of Carroll (2003)'s findings on the volume of news, we weighed tokens both within and over time in the matrix.[4] This weighting assumes that the volume of a given topic acts as a signal to agents to become more rational, increasing absorption of the specific economic topic, thereby adjusting their expectations closer to the actual outcome. Assume that the raw count for a given term to be defined as $f_{t,j}$ and the frequency as $f^{freq}_{t,j} = f_{t,j} / \sum_{j` \in t} f_{t,j`}$, where t is a corpus of text in time and j the token. The relative frequency of the term across time is then normalised by dividing it by the maximum frequency for a given training period T_{n-m}:

[3] See section 2.2 for an overview of the data, cleaning, and transformations.

[4] As opposed to the usual term-frequency, inverse document frequency (tf-idf) which tries to identify unique words.

$$f_{t,j}^{rel} = \frac{f_{t,j}^{freq}}{\max\{j`, \in T_{n-m}\} f_{t,j'}^{freq}} \tag{5.1}$$

where m is specified so as to divide the time series corpus into a training (in-sample) and test (out-of-sample) set.[5] Setting m so that the data is split on January 2012, the training sample contained 41 observations, while our out-of-sample period contained 21 quarters. The same transformation needed to be applied to the test portion of the data, T_{n+m} using the maximum frequency from the test period as the denominator. This was to ensure that no information leakage occurred between the train and test periods.

The final step before the selection procedure could be carried out was to apply near-zero variance analysis to the features. In many cases, having predictor variables with low cardinality (also known as zero variance) will decrease model performance. This is most evident when estimating linear regressions where numerical problems occur if any of the estimators are near-zero variance (NZV) predictors. In the case of text analysis, this problem is exacerbated, as some tokens might only occur in one quarter of the text sample. Given that the aim of the chapter is to identify a domain-specific dictionary, if NZV predictors were included, the RF model could overfit using these single occurrence tokens with a specific level of the outcome variable. This would result in the algorithm including words in the final dictionary that have no relevance to the outcome or, in our case, absolutely no economic connotation. Identifying and removing NZV predictors also plays a significant role in reducing the feature space of tokens. When working with bi- and unigram tokens, the feature space explodes into the millions, while most of the features have no informational value. To identify predictors with NZV characteristics, two properties can be examined (Kuhn and Johnson 2013).

The first property takes into account the percentage of unique values. The higher the percentage of unique values, the less information value the predictor has, and it will not generalise well when we apply the dictionary out of sample. There is no correct value or proportion which optimises the distinction between NZV

[5] The choice of m will influence the weighting outcome as the token weight remains anchored to its max within the training sample.

and non-NZV. We used an arbitrary unique value cut-off of 20 as it corresponds to a token needing to be unique in half of the training sample period. The second criterion on which a token is examined is the skewness of the frequency distribution of the variable. If the most frequent occurance is a much larger factor of the second most frequent token, the predictor might be highly skewed and the frequency of the variable imbalanced. Kuhn and Johnson (2013) suggests that if the most frequent value is a factor of 20 of the second most frequent value, it should be discarded. This is not a feasible number as we are working with quarterly data and a training sample of only 41 observations. We decided to apply a much stricter filtering rule of a frequency cut-off factor of 1.5.

Both of these criteria are used together to flag variables that could be potential NZV predictors. After applying this filtering process, the number of possible tokens for the dictionary dropped significantly from 14.6 million tokens to 55 000. This step could potentially be used to further scale down the number of tokens under consideration in order to lessen the computational burden, although it was not further researched in the scope of this dissertation.

5.2.2 Feature selection through RFE

Having filtered out the tokens with little to no informational value, the feature set still contained 55 000 tokens from which to generate a domain-specific dictionary. From a practical viewpoint, a dictionary with very few tokens that is able to capture the underlying trend in the economic indicator is beneficial due to its ease of dissection. If a dictionary only has a few hundred words, it is easier to understand what is causing the shift in the constructed sentiment index. In terms of a statistical property, fewer tokens introduce less noise (or complexity) into the modelling process, which could, in turn, negatively affect the desired outcome of a good fit. Certain models such as tree- and rule-based models have a natural resistance to non-informative predictors purely due to their mathematical constructs.

To select the final tokens, we employed a wrapper method known as recursive feature elimination (RFE) described in Guyon and Elisseeff (2003). This backwards selection algorithm is the main procedure that is used to create the domain-specific dictionary. It starts off by estimating a model over the whole feature space, ranking the importance of each variable by some measure. In the case of Random Forests, this is done through importance criterion such as impurity, corrected impurity, or permutation. Once the initial fit has been estimated, only the S_i most important variables (tokens) are kept and a new model is estimated from these remaining predictors. This process is continued for the specified subset of predictor variables specified by the *user* over either resampling or time-slices as in the case of time series.

Algorithm 1: Backward selection via the recursive feature elimination algorithm

```
1  for each timeslice iteration do
2      Partition data into training and hold out set via timeslices;
3      Tune/train the model on the training set using all $P$ predictors;
4      Calculate model performance;
5      Calculate variable importance or rankings;
6      for each subset size $S_i$,   $i = 1, \dots S$ do
7          Keep the $S_i$ most important variables;
8          Tune/train the model on $S_i$ predictors;
9          Calculate model performance;
10         Recalculate variable importance or rankings;
11     Calculate the performance profile over the $S_i$ using the held-back
        sample;
12     Determine the appropriate number of predictors (Set of $S_i$ associated
        with best performance);
13     Fit the final model based on the optimal $S_i$;
```

where this dissertation selected $S = \{100, 150, 200, \dots, 3500\}$. This results in the domain-specific dictionaries containing between 100 and 3500 tokens (multiples of 50). This selection was done, as searching over the whole feature space would be very time consuming.

The chosen model used in steps of 3, 8, and 14 of the RFE algorithm (1) is the RANdom forest GEneRator (Ranger) algorithm described in Wright and Ziegler (2015). The Ranger RF algorithm is a highly optimised Random Forest implementation in C++ with a focus on the analysis of highly dimensional data.

When we use RFs for regression, the procedure is making use of bootstrap aggregation or bagging. This fits the regression tree to many bootstrap-sampled versions of the training data and aggregates the estimations for the final result. Trees are excellent candidate algorithms for the concept of bagging as they have the ability to capture complex interactions in the structures of the data, while having relatively low bias if grown significantly deep. The general Random Forest algorithm described in Breiman (2001a) can be characterised as follows:[6]

Algorithm 2: Random forest for regression

```
for $\underline{b = 1 \text{ to } B}$ do
    Draw bootstrap sample $\mathbf{Z}^*$ of size $N$ from training data;
    Grow a RF tree $T_b$ to the bootstrapped data by recursively repeating
      following steps for each terminal node of the tree until minimum node
      size $n_min$ is reached;
    while $\underline{n_{min}! = min}$ do
        Select $m$ variables at random from $p$ variables;
        Pick the best variable to split among $m$;
        Split node into 2 daughter nodes;
Output ensemble of tree $T_{b1}^{B}$ Make prediction $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
```

The choice of variable importance in algorithm (1) plays a significant role in which variables are kept in S_i. Although several variants of importance criterion exist, it has been shown that by using the well known Gini impurity variable importance measure, the selection of features are biased towards features with more categories or continuous variables (Strobl et al. 2007). Given the known bias of the aforementioned, for the RFE procedure we use the permutation accuracy importance measure in its estimation step. The procedure selects a strong predictor, X_i, of the outcome and randomly permutes it. This results in its original relationship with the outcome variable no longer being retained. The permuted variable is then used in combination with the remaining unpermutated variables to predict Y. If the prediction accuracy decreases substantially, the original X_i can be deemed to be an important variable. Besides overcoming the bias associated with the univariate screening methods, the permutation test not only tests each predictor individually, but also takes into account multivariate interactions with the other predictor variables.

Next, the polarity of a given word is estimated. Once the final set of tokens has been selected as part of the RFE procedure, a simple linear model is used:

[6] For an extensive overview of tree-based methods see Friedman, Hastie, and Tibshirani (2001).

$$Y = \beta X_i + \varepsilon \tag{5.2}$$

where Y is the outcome variable and X_i is the remaining token $i \in S_i$. If the resulting coefficient β_i is greater than zero, the variable is deemed to have a positive relationship and vice versa. This estimation occurs for all $\mathbf{X}$ resulting in a final word list.

The whole algorithm from start to finish can succintly be described as:

Algorithm 3: Creating domain-specific dictionary and sentiment index

```
Tokenise time-series text into $n$-grams;
Convert into tokens into document feature matrix with $T$ as document;
Apply time-series weighting to train design matrix as in eq (5.1);
Apply weighting to out of sample;
Use near-zero variance for feature elimination;
Select $X_i$ features using RFE procedure for training data $T$;
for each selected $X_i$ do
    Fit linear regression $Y = \beta X_i + \varepsilon$;
    if $\beta > 0$ then
        Polarity $= Pos_i$;
    else
    if $\beta < 0$ then
        Polarity $= Neg_i$;
    else
        Polarity $= Neutral_i$;
Construct sentiment index $I_{it} = \frac{\sum Pos_{it} - \sum Neg_{it}}{(\sum Pos_{it} + \sum Neg_{it})}$ $\forall$ $T$ from $\mathbf{X}$;
```

5.2.3 Practical considerations

In creating a domain-specific dictionary and its respective sentiment index, several different specifications were evaluated. The experimental design is aimed at evaluating four different levers that any researcher will be faced with when applying the procedure described in algorithm (3). This resulted in having to estimate 16 different models for each of the indicators.

The first adjustable lever in automatic dictionary creation is the restriction placed to determine which tokens are eligible. Initially, the procedure is run without any restrictions on the 55 000 tokens remaining after the NZV filtering. The resulting model will be referred to as the *unrestricted* model for the remainder of the chapter. Afterwards, we placed a restriction on the words by saying that a token needs to be in the Loughran dictionary: $(W_i) \in W_L$ where W_i represents a token. For unigram tokens, this straightforward, but for the bigrams, a different approach is considered: $(W_{i1} \in W_L)$ OR $(W_{i2} \in W_L)$. The bigram is broken up into two separate unigrams and if either of them is found in the Loughran dictionary, then the bigram remains as a token. Thus, the Loughran dictionary is used as a prior in the generating process.[7] This restricted specification is referred to as the *Loughran prior* model when the results are discussed. By applying this restriction on the tokens, the feature space decreases from 55 000 to 5600.

The second choice the user has to make is the number of trees to use in the estimation step. The general rule is that more trees are required for stable variable importance estimates, but it does increase the estimation and prediction time. To test the influence of the number of trees selected for the procedure, we run both a 500- and 1000-tree variant of the model.

Next, the tuning parameter that decides on the number of variables that can be selected at each split of the tree node needs to be given: *mtry*. In a study conducted by Genuer, Poggi, and Tuleau-Malot (2010), the authors examined what influence the *mtry* parameter had on the variable importance measure. They concluded that using a large mtry leads to much higher magnitudes on the variable importance measure, which, if using the Gini method, could bias the outcome. GrÃűmping (2009) found, in turn, that a high *mtry* parameter gives lower importance to weak predictor variables in high-dimensional datasets, especially when the trees are deep. If left to the default, very shallow trees will be built and the importance of each token will be almost equally distributed. The influential impact of the *mtry* tuning parameter was tested using the following specification: high $= \sqrt{p}$ and low $= \sqrt{p}/3$, where p is the number of tokens considered. The

[7] The use of the word prior is not to be confused with the prior used in Bayesian methods, although it acts in a similar way to giving the algorithm a starting point for the dictionary creation.

default recommendation for *mtry* in a regression problem is $\sqrt{p}/3$, but given the high-dimensional nature of text tokens, it is possible that having a higher *mtry* could lead to better filtering of important tokens in the variable importance step of the procedure. To achieve an even deeper tree, the minimum node size was decreased. The minimum node size was set to 1.

The last consideration taken into account is the time-dependence nature of the outcome variable. To account for the autocorrelation inherent in time-series data, a separate specification was introduced where the design matrix included the previous period's outcome in levels, as well as the percentage change between periods. This means that the model had to its availability the choice of conditioning its selection of tokens on the time-series characteristic of the outcome variable.

5.3 Evaluation of domain-specific dictionaries and sentiment indices

This section starts by discussing and comparing the different dictionaries generated using the procedure presented in algorithm (3). We then proceed to evaluate the sentiment indices constructed using the generated dictionaries for each of the six important sentiment indicators produced by the BER and the SARB namely: Business Confidence (BCI), Building Confidence (Build), Consumer confidence (CCI), Civil Confidence (Civil), Purchasing Managers's index (PMI) and the SARB Leading Business Cycle indicator (Leading).

5.3.1 Generated dictionaries

Table 5.1 shows all of the model specifications, the number of tokens, and the proportion of positive and negative words for each resulting dictionary. We can see that the unrestricted models generated much larger dictionaries for the BCI, Build and Civil confidence indices, while for CCI and PMI, all specifications generated large dictionaries. This would suggest that the models that performed the best in terms of cross-validation were more complex models consisting of a large number of tokens as predictors. Another observation from the table is the low number of tokens selected when we used the Loughran dictionary as a prior. A large proportion of these dictionaries only contain 100 tokens, the minimum set as specified in the RFE procedure.

Table 5.1: List of specifications under evaluation and the number of tokens that were selected by the RFE procedure. The number in the bracket shows the proportion positive and negative words (positive, negative).

Specification	Type	Inc lag	Tree size	Mtry size	Bci	Build	Cci
A	loughran	FALSE	500.00	high	100 (27%,72%)	950 (31%,68%)	3500 (29%,69%)
A	unrestricted	FALSE	500.00	high	3300 (43%,56%)	1800 (46.3%,51.8%)	3500 (49.40%,50.14%)
B	loughran	TRUE	500.00	high	100 (25%,73%)	200 (30%,67%)	3400 (29%,69%)
B	unrestricted	TRUE	500.00	high	2850 (44%,56%)	3100 (44.9%,53.9%)	3350 (49.16%,49.97%)
C	loughran	FALSE	1000.00	high	100 (26%,73%)	200 (29%,68%)	3450 (29%,69%)
C	unrestricted	FALSE	1000.00	high	3000 (42%,57%)	3300 (43%,56%)	3250 (49.75%,49.63%)
D	loughran	TRUE	1000.00	high	100 (26%,72%)	100 (27%,70%)	3250 (30%,69%)
D	unrestricted	TRUE	1000.00	high	2100 (43%,56%)	3300 (44%,55%)	3400 (49.82%,49.41%)
E	loughran	FALSE	500.00	low	100 (25%,73%)	100 (25%,72%)	2500 (32%,66%)
E	unrestricted	FALSE	500.00	low	2550 (44%,55%)	2550 (43%,55%)	3400 (51.4%,48.1%)
F	loughran	TRUE	500.00	low	100 (22%,76%)	100 (24%,72%)	1550 (36%,63%)
F	unrestricted	TRUE	500.00	low	3050 (44%,55%)	2100 (45.6%,52.9%)	2700 (53.3%,46.0%)
G	loughran	FALSE	1000.00	low	100 (21%,78%)	100 (26%,71%)	3250 (30%,68%)
G	unrestricted	FALSE	1000.00	low	1750 (42%,56%)	2200 (45.9%,52.3%)	2750 (54.5%,44.9%)
H	loughran	TRUE	1000.00	low	100 (23%,75%)	100 (25%,72%)	3100 (31%,67%)
H	unrestricted	TRUE	1000.00	low	3250 (44%,55%)	2650 (44.8%,53.5%)	3500 (52.7%,46.7%)

Specification	Type	Inc lag	Tree size	Mtry size	Civil	Leading	Pmi
A	loughran	FALSE	500.00	high	100 (34%,60%)	550 (43%,57%)	3050 (49.77%,48.79%)
A	unrestricted	FALSE	500.00	high	3250 (47.4%,51.0%)	100 (76%,23%)	3400 (35%,64%)
B	loughran	TRUE	500.00	high	100 (34%,59%)	150 (53.3%,46.0%)	3450 (52.0%,46.3%)
B	unrestricted	TRUE	500.00	high	2700 (47.5%,50.8%)	100 (74%,24%)	3300 (34%,65%)
C	loughran	FALSE	1000.00	high	100 (33%,62%)	150 (54.0%,46.0%)	3450 (51.3%,47.0%)
C	unrestricted	FALSE	1000.00	high	3250 (45.6%,52.6%)	150 (74%,25%)	3350 (33%,65%)
D	loughran	TRUE	1000.00	high	100 (33%,62%)	100 (57%,42%)	800 (25%,74%)
D	unrestricted	TRUE	1000.00	high	3450 (44.9%,53.3%)	100 (74%,24%)	2750 (30%,68%)
E	loughran	FALSE	500.00	low	100 (34%,61%)	3450 (31%,67%)	2250 (43%,56%)
E	unrestricted	FALSE	500.00	low	3200 (47.9%,50.7%)	150 (71%,27%)	1350 (24%,74%)
F	loughran	TRUE	500.00	low	100 (34%,58%)	100 (54.0%,45.0%)	350 (11.4%,86.3%)
F	unrestricted	TRUE	500.00	low	3500 (46.7%,52.2%)	150 (75%,23%)	1500 (26%,73%)
G	loughran	FALSE	1000.00	low	100 (33%,61%)	250 (50.4%,49.2%)	1850 (37%,62%)
G	unrestricted	FALSE	1000.00	low	3350 (46.5%,51.5%)	200 (72%,26%)	3300 (30%,68%)
H	loughran	TRUE	1000.00	low	100 (32%,60%)	100 (53.0%,46.0%)	450 (16%,82%)
H	unrestricted	TRUE	1000.00	low	2550 (47.1%,50.7%)	200 (76%,23%)	3400 (32%,67%)

To measure the overlap for a given outcome and its various dictionaries, we looked at how many of the words occur in all dictionaries. The maximum number of words is bound by the smallest dictionary that is generated. For the Loughran prior specifications, the number of similar tokens among the different dictionaries is quite high. For instance, in the case of the BCI, 74 tokens were similar in eight of the generated dictionaries from different specifications, all of whom only had 100 tokens each.

Table 5.2: Breakdown of the similarities among the different dictionaries generated from the various specifications.

Sentiment	Type	Bci	Build	Cci	Civil	Leading	Pmi
Distinct Tokens	loughran	127	951	4708	133	3451	4567
Overlapping tokens	loughran	74	82	1346	77	83	330
negative	loughran	53 (72%)	59 (72%)	846 (63%)	45 (58%)	36 (43%)	287 (87%)
neutral	loughran	1 (1%)	2 (2%)	10 (1%)	3 (4%)	0	5 (2%)
positive	loughran	20 (27%)	21 (26%)	490 (36%)	29 (38%)	47 (57%)	38 (12%)
Distinct Tokens	unrestricted	6498	6045	7422	7242	245	7397
Overlapping tokens	unrestricted	748	722	903	870	64	545
negative	unrestricted	402 (53.7%)	359 (49.72%)	371 (41%)	401 (46.1%)	14 (22%)	437 (80%)
neutral	unrestricted	6 (0.8%)	9 (1.25%)	6 (1%)	14 (1.6%)	1 (2%)	6 (1%)
positive	unrestricted	340 (45.5%)	354 (49.03%)	526 (58%)	455 (52.3%)	49 (77%)	102 (19%)

This indicates that although the choice of model specification has a large influence on the final number of distinct tokens selected by all the different specifications, there are core tokens that occur in all dictionaries. To gain some perspective on the bi-gram tokens that were selected for the dictionaries, we visually represent key words.

Using network graphs, figures (5.1) to (5.3) show the network among bi-gram text for each of the outcome variables. To construct the network graph, all dictionaries (for a given outcome) were collapsed and the three bi-grams with the highest degree of *betweenness* were kept.[8] The betweenness centrality measure indicates which tokens have a high flow of information passing through them and is principally based on shortest path algorithms. Thus, the tokens with a high degree of betweenness form central concepts key to the various dictionaries.

[8] See Freeman (1977) for the seminal work on betweenness centrality.

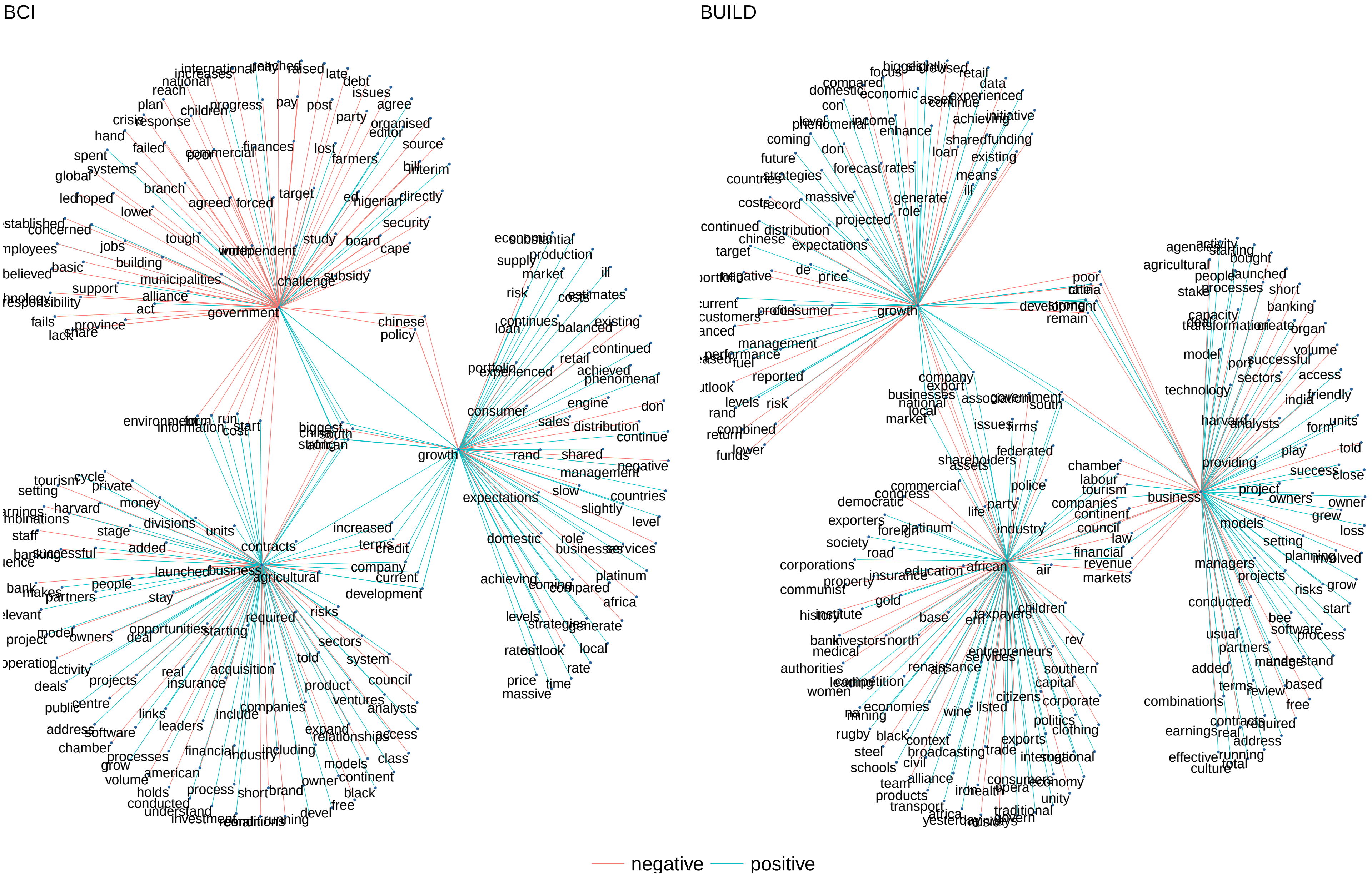


Figure 5.1: Network graph based off of bi-gram tokens filtered on the top three highest measures of betweenness centrality (BCI/Build).

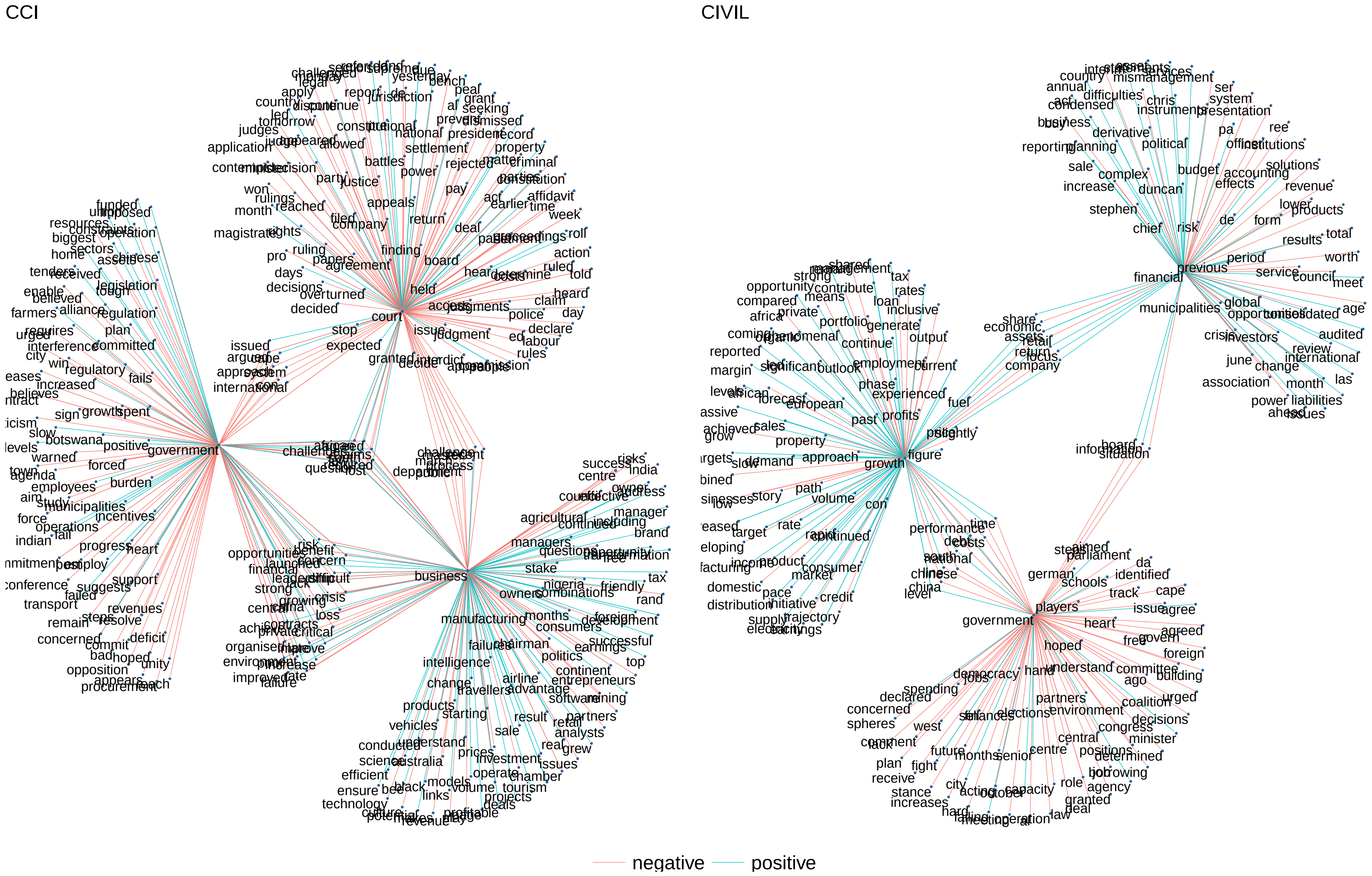


Figure 5.2: Network graph based off of bi-gram tokens filtered on the top three highest measures of betweenness centrality (CCI/Civil).

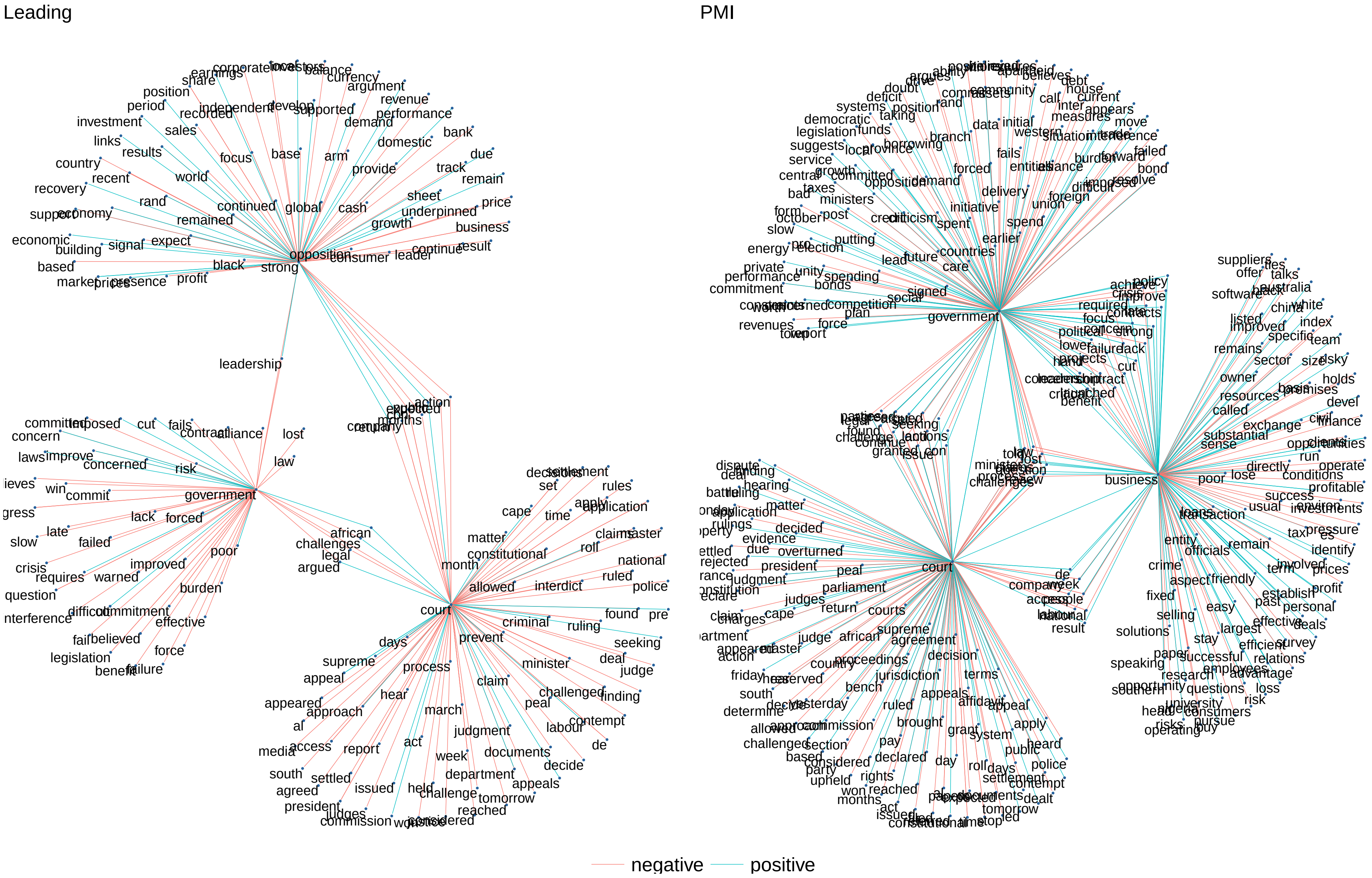


Figure 5.3: Network graph based off of bi-gram tokens filtered on the top three highest measures of betweenness centrality (Leading/PMI).

Figures (5.1) to (5.3) show that for government, the tokens tend to have a negative polarity, while growth has more positive tokens connected to it. This could be due to the bias found within the media that tends to report negative news relating to government activities more eagerly than positive news. Table 5.3 gives the rank of betweenness for the top ten tokens as per each of the different outcomes. The table contains 20 unique tokens that were isolated. Tokens that were selected across all the outcomes include 'business', 'company', 'government', and 'market'. Although the token 'growth' does not appear in the PMI-generated dictionary, it is considered to be important for all other measures.

Table 5.3: Rank of the top ten words as per the betweenness centrality measure for each of the outcomes. In total, 20 unique tokens were isolated and of these 'business', 'company', 'government', and 'market' occurred in all outcome dictionaries.

token	BCI	BUILD	CCI	CIVIL	Leading	PMI
african	5	3	6	-	-	-
business	2	1	3	4	7	3
company	7	6	9	6	5	5
court	-	-	1	-	1	1
economic	-	10	-	7	-	-
financial	10	7	-	2	-	10
global	8	-	-	-	-	-
government	3	4	2	3	2	2
growth	1	2	4	1	8	-
investment	-	-	-	10	-	-
law	-	-	10	-	6	-
market	4	5	7	5	9	8
national	9	-	-	8	-	-
people	-	-	-	-	-	6
poor	-	-	-	-	10	-
risk	-	-	5	-	4	4
share	-	-	-	9	-	-
south	6	9	-	-	-	-
strong	-	8	8	-	3	9
time	-	-	-	-	-	7

We can see that for the BCI, CIVIL, and BUILD indices, the term growth had the highest (or second) highest measure of centrality, while for the CCI, SARB leading indicator, and PMI, the token 'court' had the highest value.

Examining figures (5.1) to (5.3) and table 5.3 confirms that the algorithm's variable selection method is isolating words that have meaning within the broader context of economic sentiment. Although these tokens or their polarity do not necessarily equate to their linguistic orientation, they are capturing some form of sentiment towards the economy, business, government, growth, and the markets. To evaluate how well these automated dictionaries capture sentiment, we constructed sentiment indices and analysed out-of-sample fits between the various confidence measures and the constructed sentiment index.

5.3.2 Constructing and evaluating sentiment indices

This section deals with the construction and evaluation of various sentiment indices using the generated dictionaries. These indices are compared to a sentiment index that is constructed using the well-known Loughran & McDonald dictionary as a baseline. To evaluate these indices, we use the root mean squared error between the constructed indices and the respective confidence measure as an indication of fit:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(s_t - y_t)^2}{T}} \tag{5.3}$$

Where s_t is the sentiment index at t and y_t confidence measure.

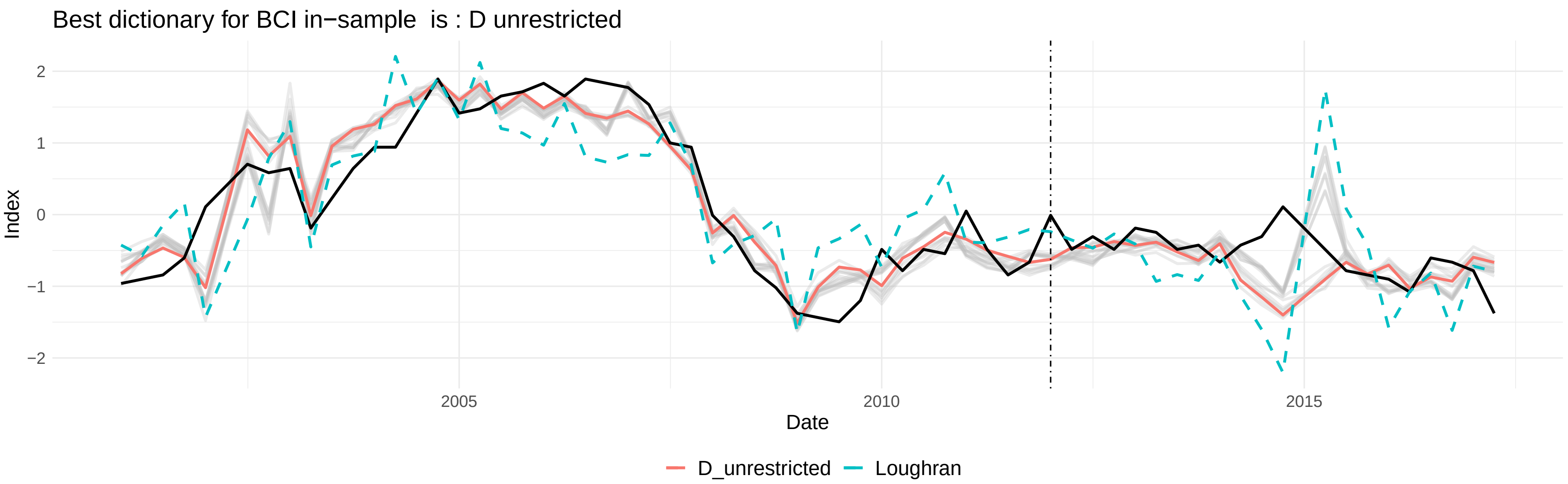
Following Odendaal, Reid, and Kirsten (2018), we constructed the indices using a net score. We identified the positive and negative words for all article $Nt \quad \forall \quad T$ using the generated word lists (dictionary). Conducting a simple word count that consists of the positive plus negative words, a sentiment index was created. We normalised the count so that it reflects the relative proportion of positive and negative words within a period:

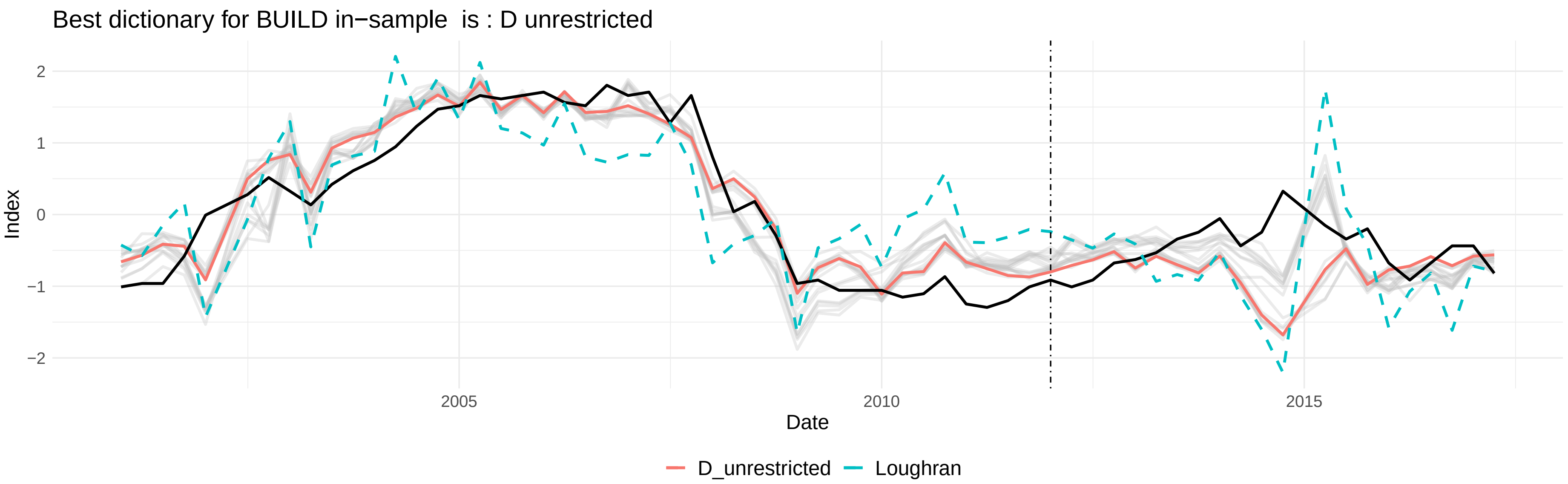
$$Pos_t = \frac{PositiveWords}{PositiveWords + NegativeWords} \quad Neg_t = \frac{NegativeWords}{PositiveWords + NegativeWords} \tag{5.4}$$

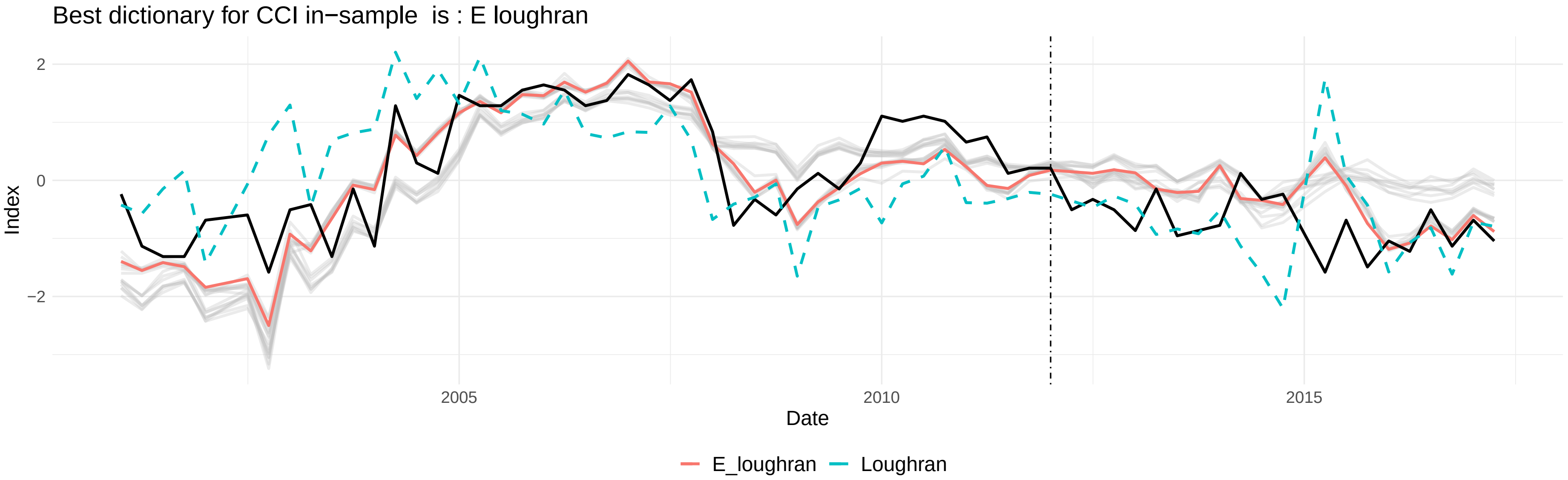
The overall sentiment index for a given time period t can then be defined as:

$$S_t = Pos_t - Neg_t \tag{5.5}$$

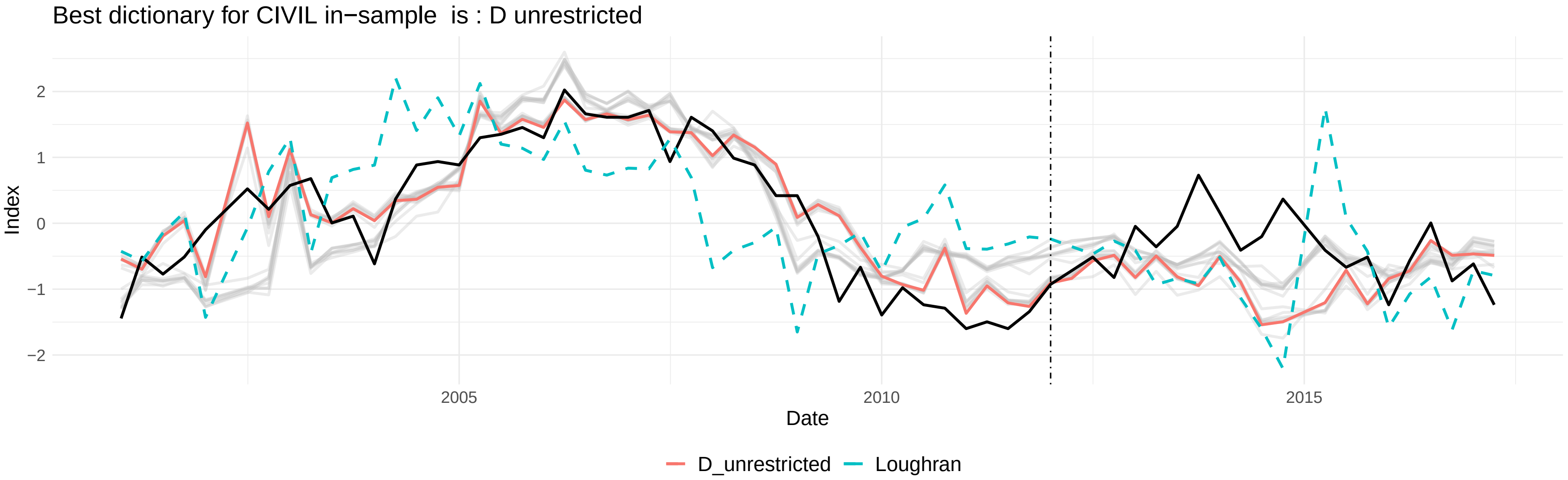
The resulting index is the net balance of positive and negative words within a quarter. Figure (5.4) shows the constructed indices, the sentiment index constructed using the baseline dictionary as well as the confidence measure. For each of the series, the dictionary with the lowest RMSE in-sample is highlighted. For three of the indices, BCI, BUILD, and CIVIL, the best RMSE was achieved using specification D. This entailed including the lag (and change) of the underlying series in the predictor set, building 1000 trees, not using Loughran as a prior, and finally, setting the *mtry* tuning parameter to $\sqrt{p}$. The second most used specification was E (CCI and SARB Leading) that uses the Loughran dictionary as a prior, does not include properties of the outcome series into the predictor set, only builds 500 trees, and uses a low number of variables to split on ($mtry = \sqrt{p}/3$). For the PMI, the sentiment index that had the lowest in-sample RMSE was A, the Loughran prior specification. This specification is identical to the specification that had the lowest RMSE for the CCI and the SARB leading indicator, but it had the *mtry* parameter as high $= \sqrt{p}$.


Best dictionary for BCI in−sample is : D unrestricted
Index
2
1
0
−1
−2
2005
2010
2015
Date
D_unrestricted
Loughran


Best dictionary for BUILD in−sample is : D unrestricted
Index
2
1
0
−1
−2
2005
2010
2015
Date
D_unrestricted
Loughran



Best dictionary for CCI in−sample is : E loughran



Best dictionary for CIVIL in−sample is : D unrestricted

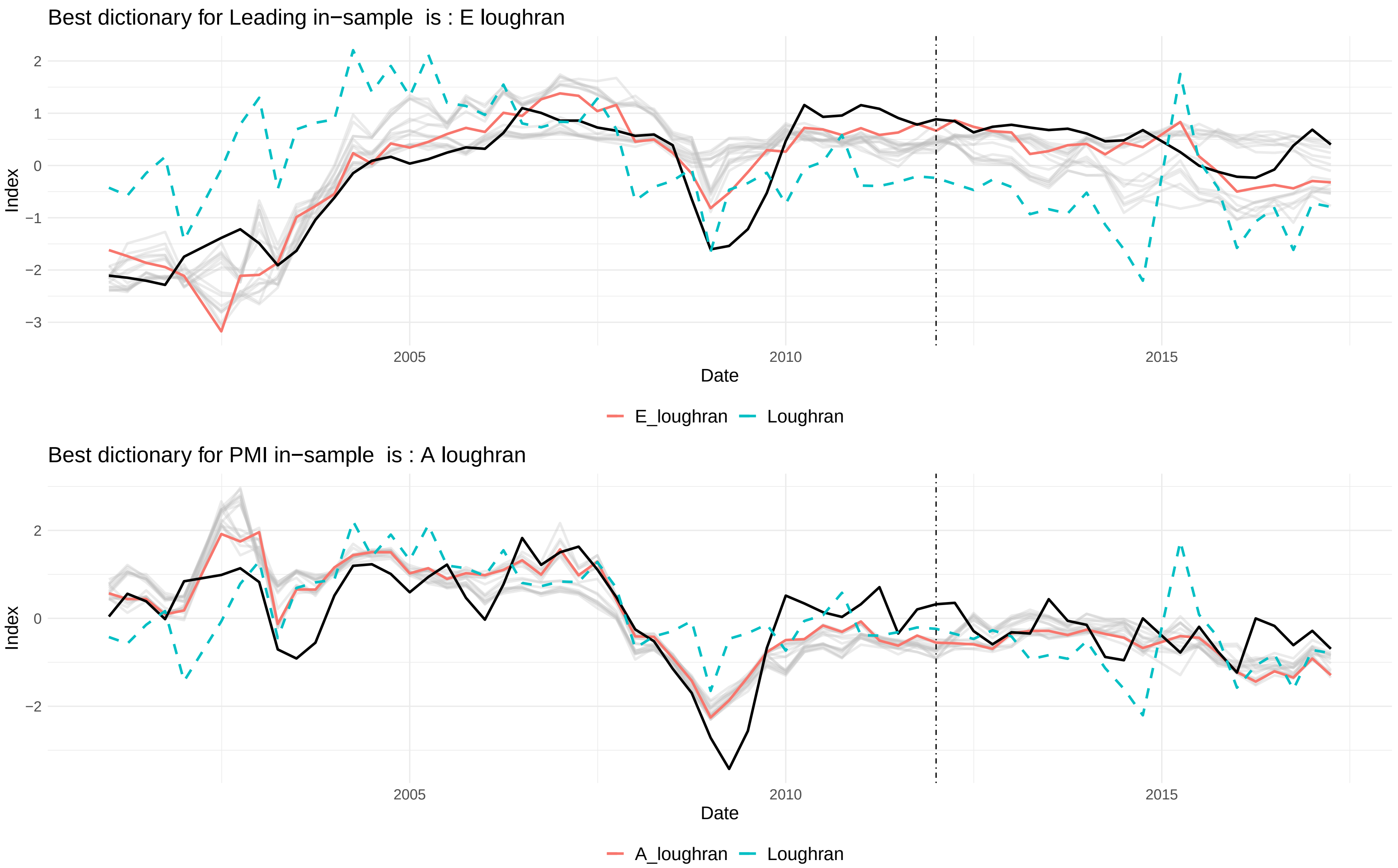


Figure 5.4: Figure illustrating the fit of the constructed sentiment indices with their respective confidence outcome measure.

To evaluate the underlying relationship between the generated series and the confidence measure, we calculated the cross-correlation between the series out-of-sample. Cross-correlation can be defined as a measure of the similarity between two series as a function of the lag of the predictor relative to the outcome. We conducted this test in order to assess whether generated sentiment indices perhaps lag or lead the confidence measure. We used the specification which had the lowest RMSE in-sample and the traditional Loughran McDonald dictionary and calculated the cross-correlation between the series and the outcome. The series were all tested for stationarity before estimating the cross-correlation and found to be non-stationary, so all series were made stationary using first log difference. The series were tested using a maximum lag of up to four quarters.

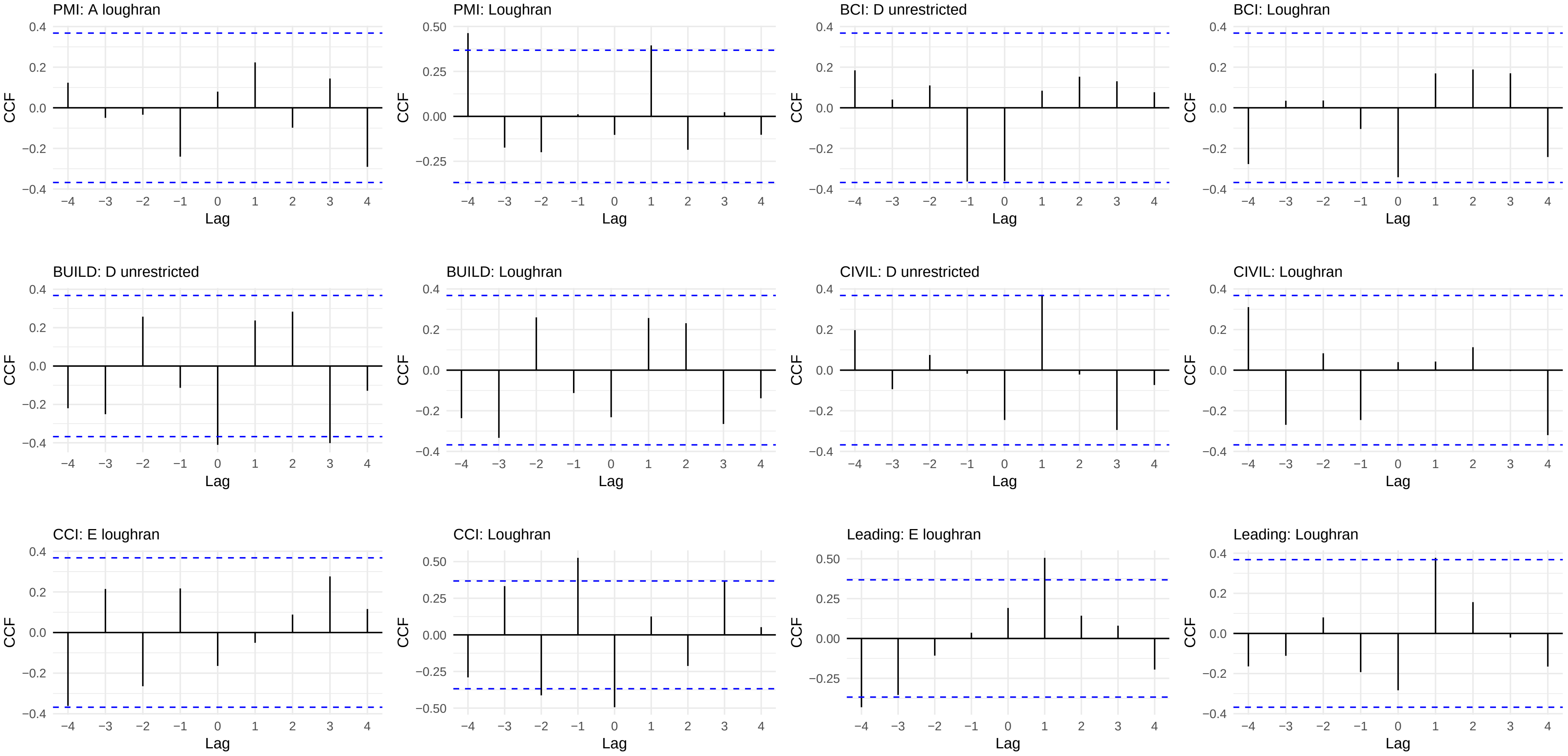


Figure 5.5: Cross-correlation between the constructed sentiment indices with their respective confidence outcome measure. The statistical significance of the correlation between the series is tested at the 10% level and is indicated by the dotted lines.

Figure 5.5 shows how the sentiment indices correlate with changes observed in the CCI, leading indicator, and PMI. In the case of the CCI, the correlation is negative, while the correlation is positive for the leading indicator and the PMI. For the generated dictionaries, fewer of the correlations were statistically significant. A contemporaneous negative correlation is seen between the BUILD confidence index and specification *D unrestricted.* The other significant relationship is between the leading indicator and *E unrestricted.* Although correlation is a way to understand the dynamic relationship between two series, it does not tell us how well the sentiment indices fit the original confidence indicators. To evaluate this, we looked at the out-of-sample fits.

Given the nature of the algorithm and the possibility that it is overfitting the series in-sample, we could not compare the sentiment indices generated from the newly generated dictionaries with the well-known Loughran dictionary. We turned to an out-of-sample RSME evaluation of the various series.[9] Table (5.4) contains the RMSE measures as well as the percentage difference of each specification to the specification that achieved the lowest RMSE in-sample. Using BCI as an example, the specification that achieved the lowest RMSE in-sample was *D unrestricted.* This same specification achieved an RMSE of 0.475 out-of-sample. Using this number as the base, we compare the RMSE of the other specifications with 0.475 to evaluate the best in-sample specification's performance out-of-sample. On average, the domain-specific dictionaries generated for the BCI only differed by -2% out-of-sample, while for the SARB leading indicator, the difference was much larger at 37.6%.

[9] The authors also calculated various other error measures which are available on request: ME, MAE, MPE, and MAPE.

Table 5.4: The RMSE measure as calculated by comparing the generated series and the actual indices based on 20 quarterly out-of-sample observations from 2012 to 2017. The values represent the RMSE measures per specification as well the percentage difference in error of each specification with the specification that had the lowest RMSE in-sample (in brackets).

Specification	Type	Inc lag	Tree size	Mtry size	Bci	Build	Cci
A	loughran	FALSE	500	high	0.509 (7.2%)	0.493 (-19.0%)	0.693 (4.0%)
A	unrestricted	FALSE	500	high	0.430 (-9.6%)	0.581 (-4.6%)	0.893 (34.0%)
B	loughran	TRUE	500	high	0.482 (1.4%)	0.437 (-28.2%)	0.683 (2.4%)
B	unrestricted	TRUE	500	high	0.500 (5.2%)	0.627 (3.0%)	0.840 (25.8%)
C	loughran	FALSE	1000	high	0.445 (-6.4%)	0.483 (-20.8%)	0.662 (-0.8%)
C	unrestricted	FALSE	1000	high	0.500 (5.2%)	0.627 (3.0%)	0.883 (32.4%)
D	loughran	TRUE	1000	high	0.454 (-4.4%)	0.455 (-25.2%)	0.690 (3.4%)
D	unrestricted	TRUE	1000	high	0.475 (-)	0.609 (-)	0.735 (10.2%)
E	loughran	FALSE	500	low	0.471 (-0.8%)	0.454 (-25.4%)	0.667 (-)
E	unrestricted	FALSE	500	low	0.437 (-8.0%)	0.644 (5.6%)	0.884 (32.4%)
F	loughran	TRUE	500	low	0.428 (-10.0%)	0.446 (-26.8%)	0.676 (1.2%)
F	unrestricted	TRUE	500	low	0.452 (-4.8%)	0.637 (4.6%)	0.942 (41.2%)
G	loughran	FALSE	1000	low	0.435 (-8.4%)	0.436 (-28.4%)	0.711 (6.6%)
G	unrestricted	FALSE	1000	low	0.482 (1.4%)	0.603 (-1.0%)	0.872 (30.8%)
H	loughran	TRUE	1000	low	0.495 (4.2%)	0.444 (-27.0%)	0.675 (1.2%)
H	unrestricted	TRUE	1000	low	0.451 (-5.0%)	0.610 (0.2%)	0.788 (18.2%)
loughran	-	-	-	-	0.880 (85.2%)	0.907 (49.0%)	1.019 (52.8%)
Mean diff	-	-	-	-	-2.0%	-11.8%	15.2%

Specification	Type	Inc lag	Tree size	Mtry size	Civil	Leading	Pmi
A	loughran	FALSE	500	high	0.586 (-23.2%)	0.654 (59.2%)	0.616 (-)
A	unrestricted	FALSE	500	high	0.758 (-0.6%)	0.416 (1.2%)	0.553 (-10.4%)
B	loughran	TRUE	500	high	0.627 (-17.6%)	0.756 (84.0%)	0.606 (-1.6%)
B	unrestricted	TRUE	500	high	0.701 (-8.0%)	0.394 (-4.0%)	0.502 (-18.6%)
C	loughran	FALSE	1000	high	0.583 (-23.6%)	0.737 (79.4%)	0.623 (1.2%)
C	unrestricted	FALSE	1000	high	0.764 (0.2%)	0.394 (-4.2%)	0.557 (-9.6%)
D	loughran	TRUE	1000	high	0.591 (-22.4%)	0.897 (118.2%)	0.660 (7.0%)
D	unrestricted	TRUE	1000	high	0.762 (-)	0.510 (24.0%)	0.525 (-14.8%)
E	loughran	FALSE	500	low	0.586 (-23.0%)	0.411 (-)	0.660 (7.0%)
E	unrestricted	FALSE	500	low	0.768 (0.8%)	0.416 (1.2%)	0.555 (-10.0%)
F	loughran	TRUE	500	low	0.575 (-24.6%)	0.771 (87.6%)	0.645 (4.6%)
F	unrestricted	TRUE	500	low	0.863 (13.2%)	0.439 (6.8%)	0.569 (-7.6%)
G	loughran	FALSE	1000	low	0.574 (-24.6%)	0.647 (57.6%)	0.640 (3.8%)
G	unrestricted	FALSE	1000	low	0.752 (-1.4%)	0.360 (-12.4%)	0.569 (-7.6%)
H	loughran	TRUE	1000	low	0.569 (-25.4%)	0.843 (105.2%)	0.623 (1.0%)
H	unrestricted	TRUE	1000	low	0.757 (-0.8%)	0.404 (-1.6%)	0.565 (-8.4%)
loughran	-	-	-	-	1.008 (32.2%)	1.437 (249.6%)	0.942 (52.8%)
Mean diff	-	-	-	-	-11.4%	37.6%	-4.0%

In comparison, the differences observed when comparing the Loughran dictionary's fit, is much larger. The average difference among all the outcomes is 85%. This would indicate that the domain-specific dictionaries significantly improved on the out-of-sample fit when compared to financial dictionary. The largest difference observed in RMSE measures was when comparing the fit of the SARB leading indicator. Here, the domain-specific dictionary decreased the RMSE by a factor 2.5 out-of-sample.

This led us to ask whether the specifications themselves are different and whether the choice on specification makes a significant difference on the out-of-sample performance of the sentiment index. To answer this question, we employed a non-parametric statistical method in order to test the hypothesis that the mean RMSE values of the difference specifications are the same. We used a pairwise Wilcoxon ranks sum test and compared means between group levels. The null hypothesis for the test is: true location shift is not greater than zero (i.e., the RMSE values are roughly the same).

Table (5.5) shows the p-values obtained from all the pairwise tests conducted. The insignificance of any of the values indicate that there is no statistically significant difference in the RMSE measures across the different specifications of the models. This finding highlights the fact that on average, the specification of the model does not have a significant impact on the mean RMSE measures produced out-of-sample. It has to be clarified that although this was the case for the outcomes presented in this chapter, the finding is unlikely to generalise across different datasets. Further research is needed to understand the effect of each subjective choice in the final generated dictionary.

Table 5.5: Results from pairwise Wilcoxon ranks sum test. The high p-values among all the tests indicate that there is no statistically significant difference in the RMSE measures across the different specifications.

Variables	Inc lag false	Inc lag true	Mtry size high	Mtry size low	Tree size 1000
Inc lag: TRUE	0.283	-	-	-	-
Mtry size: high	0.316	0.525	-	-	-
Mtry size: low	0.464	0.7	0.665	-	-
Tree size: 1000	0.35	0.56	0.536	0.372	-
Tree size: 500	0.426	0.669	0.633	0.464	0.597

Apart from testing the specifications among themselves, the Wilcoxon test was also employed to test whether a statistically significant difference exists between the unrestricted, Loughran prior, and the traditional Loughran sentiment indices. Figure (5.6) shows the boxplot of the pairwise Wilcoxon test between the different model types. The unrestricted and Loughran prior model types show no statistical difference between their mean RMSEs. What can be visually observed is that the Loughran prior has a smaller deviation of RMSE measures, $\sigma = 0.11$, than its unrestricted counterpart, $\sigma = 0.16$.

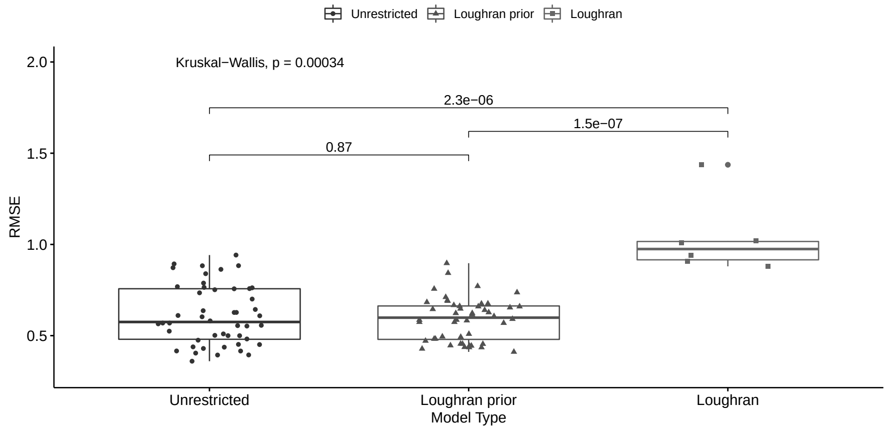


Figure 5.6: Pairwise Wilcoxon test between the different model types.

These results show that although the mean errors from the unrestricted and Loughran prior are not statistically different from one another, generated dictionaries that use a domain-specific dictionary as a prior could deliver the same results with much less noise. Using a Fligner-Killeen test, the null hypothesis of equal variance was rejected at the 5% level with a p-value of 0.0168. These findings motivate the use of a domain-specific dictionary, that has been proven effective through human verification, as a starting point when using machine learning algorithms to generate dictionaries for sentiment analysis.

5.4 Conclusion

In this chapter, machine learning was presented as one way to create a domain-specific dictionary. This approach has been shown to be able to create sentiment lexicons tailored to a specific need. These dictionaries, being less subjective, are more easily tested and replicable. Specific word lists also contribute to transparency since it is easy for other researchers to replicate the results.

Tokens (words) were statistically selected using recursive feature elimination to generate a domain-specific dictionary from a corpus of text. The machine-generated dictionaries consist of both uni-grams and bi-grams. The inclusion of bi-grams add context to the sentiment dictionary, ensuring that the tokens themselves are less ambiguous. To try and understand the bi-grams in the dictionaries, we visually show that they contain central ideas (topics) that help in capturing sentiment for each of the confidence measures. These ideas ranged from issues in government, business, and growth, as well as what is happening in the markets. This concentration of tokens around key concepts substantiates the hypothesis of Sims (2003)'s rational inattention model. Having to focus on the economy and its intricacies has an opportunity cost connected to it; in turn, the agents would rather process information on economic activity through concentrated sources, such as the media (and expert reporting), where key information on business, markets, government, etc. can be more easily digested. This distillation of key information can be tracked in a generalised construct known as sentiment indices.

The results of this chapter show that the indices constructed from machine-generated dictionaries have a better fit with the multiple indicators investigated compared to the sentiment index constructed from a commonly used financial dictionary. These domain-specific sentiment indices also show a significantly lower root mean squared error (RMSE) in a five-year holdout sample period from 2012 to 2017. The largest improvement was observed for the leading indicator, where the domain-specific dictionary improved the fit by a factor of 2.5.

These results support the case for domain-specific dictionaries being able to pick up nuances found within domain-specific topic news. The results, however, suggest that having a manually generated dictionary act as a prior narrows the tokens[10] that the Random Forest has to search over, while maintaining the same lower RMSE out-of-sample as an unrestricted model with all tokens. Employing a manually generated dictionary such as Loughran and McDonald (2011)'s also decreases the computational burden on the pre-processing and estimation of the dictionaries. Another finding of the chapter relates to practical considerations when implementing the Random Forest algorithm tuning parameters. In the case of this thesis' data, the tuning parameters had no statistical difference in the RMSE across all different specifications. Although this was the finding for this chapter's specific research design, it is a finding that needs further research to validate and understand. A better understanding of the dynamics between the tuning parameters and the resulting dictionary could result in much more robust dictionary generation, especially in an operational setting.

[10] See 5.2 for reference.

Chapter 6

Summary and conclusion

According to Pigou (1927)'s theory of cycles, even though an economy might not be experiencing any technological regression, changes in the expectations of consumers and firms, driven by sentiment (confidence and uncertainty), can have a large impact on the real economy's business cycles (Beaudry and Portier 2004).

Since the financial crisis, substantial attention has been devoted to understanding the role of sentiment in the real economy. Economic agents react to higher degrees of uncertainty in their future income. This can have a real effect on households' precautionary savings, while also contemporaneously reducing their consumption expenditure. By capturing large shocks in economic uncertainty, household wealth, and credit consumption, and by integrating sentiment, it is possible to explain a large proportion of a household's saving and spending behaviour.

Sentiment not only affects households, but also plays a significant role in how firms decide to invest and save. General business sentiment affects decisions from both a demand and supply perspective. From the supply side, negative sentiment increases labour market attrition and the demand for new hiring decreases. On the demand side, if sentiment over the medium to long term is pessimistic, accompanied by high levels of uncertainty, fixed capital investment and purchases of durable goods might also be deferred.

Given the significant role of general sentiment in the economy, the use of confidence indicators to monitor economic developments has become indispensable during turbulent economic times, even acting in some cases as a leading indicator of movements in main economic data (Kalamara et al. 2020). Despite a lack in understanding of the causal relationship between the real economy and sentiment, the recognition of confidence indicators as important and useful tools to monitor economic development is no longer disputed. The inclusion of confidence indicators in consumer expenditure or economic growth forecasting models is not new to the literature (see Bram and Ludvigson 1997; Ludvigson 2004; Gelper and Croux 2010). In practice, however, the frequency at which these confidence indicators become available might have a substantial impact on their usefulness to policy makers and forecasters, especially during financial or economic distress. Text analysis could possibly offer an alternative or complementary representation of expectations which are available at a higher frequency.

Given the enormous increase in the volume and velocity of online text information, it is almost certain that text analysis will become a key area of research within economics. Coinciding with this increase in the availability of information has been the rapidly expanding literature of machine learning techniques which provide novel approaches to analysing large datasets. Supporting this growth of machine learning has been the equally fast-developing industry of cloud computing. Enabled by these developments, a large body of literature is emerging that is focused on capturing consumer sentiment specifically through the analysis of different text sources. To help introduce the topic of text analysis, chapter 2 uses illustrative examples to introduce readers to basic concepts within computational linguistics, and reviews the data sources presented throughout the remaining chapters.

In this dissertation, the case was made that text analysis offers a complement to traditional survey-based methods, as a way to capture sentiment. This was achieved by constructing media-based sentiment indices from a large variety of news sources for South Africa. The findings of chapter 3 support the hypothesis that news-based indices could be used to address the shortcomings found in the survey-based confidence measures such as non-response, low frequency of publication, and high associated costs. It also supports the hypothesis that domain-specific dictionaries can be constructed using semi-automated machine learning methods that help guide the construction of a representative confidence index that improve upon the commonly used Loughran dictionary.

The second contribution towards the literature is the evaluation of the adequecy of media-based indices as a complement or substitute for the CCI as a predictor of personal consumption. The predictive power of media sentiment indices are evaluated in a horse race framework alongside the CCI. The conclusion of the forecasting exercise revealed that the inclusion of media-based sentiment indices as predictors in a model can decrease forecasting errors of personal consumption expenditure. The forecasting errors not only decreased over the short forecasting horizons, but also over longer horizons of up to two years. The results substantiate the theory that news media sentiment contains information on the coincidental and forward-looking state of the economy, which is not captured in the CCI.

Collectively, the results presented in this dissertation offer some initial support for the use of text analysis in South Africa as an alternative way of capturing softer economic indicators such as economic sentiment.

6.1 Media-based sentiment indices

In South Africa, the business and consumer confidence index (BCI and CCI) are only released on a quarterly basis, limiting the information the indicator can provide on the expectations of economic agents. By using text analysis, newly constructed indices are able to capture both general and specific economic events that may have an influence on shaping consumers' and firms' views of the economy. Text-based indices also have the advantage of not having to survey agents directly on their view of the economy, but, through computational linguistics, extracting this view from business and financial reporting in the news media. These views can then be captured in a sentiment index potentially reflecting the state of the economy. Despite the advantages of news media, sampling issues and persistent bias of journalists' views could call into question the true representation of the index.

In chapter 3, I sought to investigate the feasibility of constructing online sentiment indices using various different text sources and sentiment dictionaries as candidate alternatives of the CCI. A procedure that uses a clustering framework to select the most appropriate data sources and lexicon is presented. The main objective was to investigate whether an online text-based sentiment index can offer an alternative to the survey-based approach that does not suffer from the same logistical challenges. The chapter introduced a procedure whereby multiple indices, controlling for data provider and lexicon dictionary, can be tested with respect to their ability to replicate the traditional survey-based index. Employing a time-series clustering technique and dynamic time warping as a dissimilarity measure, groups from the clustering were used to create composite indices. These composite indices use the clustering output to identify similar series to form part of a composite index. Along with the cluster-based approach for composite indices, the raw dissimilarity value was also used to select series similar to the CCI. The raw dissimilarity index was used as a counterargument to the use of composite indices that add another layer of analytical complexity. In summary, the chapter showed that it is possible to create an index using sentiment analysis to leverage large amounts of online editorial data that resemble the BER's consumer confidence index. This was shown statistically by the findings of the cointegration and Granger-causality test. The results indicated that one of the alternatives were possibly cointegrated with the CCI, while the Granger-causality tests revealed a leading relationship for two of the media-based sentiment indices.

6.2 Predictive power of text-based sentiment indices

The finding in chapter 3 that a media-based index is able to replicate the CCI to a reasonable extent implicitly assumed that the CCI was an accurate measure of sentiment. To extend the analysis of the extent to which the media based indices are able to capture valuable economic information, the thesis then turned to the ability of the indices to forecast consumer economic activity.

Chapter 4 built on the previous chapter by conducting a forecasting experiment that investigates whether media-based sentiment indices in South Africa have any predictive power in their own right. The evidence in this chapter suggests that media-based consumer confidence indices can help to improve the forecasting of personal consumption expenditure at both shorter and longer forecasting horizons.

By estimating different specifications of a Bayesian VAR where media-based sentiment indices are included and excluded, we forecasted personal consumption expenditure as well as its sub-components for forecast horizons $h = \{1, 2, 3, 4, 6, 8\}$. To evaluate the out-of-sample forecasting performance of each of the different model specifications, a point estimate RMSE and density-based scoring was employed, namely continuous ranked probability score (CRPS) and log-score (LS). In addition to the scoring evaluations, the forecasting performance was statistically analysed through a procedure called model confidence sets (MCS).

Using the median CRPS score as the primary performance measure, the results show that the inclusion of a Loughran-based sentiment index as part of the forecasting model improved forecasts for the Durable and Non-durable components of personal consumption expenditure (PCE). The same was found for total PCE. The results of the statistical MCS procedure found the same benefit for forecasting total PCE, but less convincing results for forecasting the sub-components. For Durables, the baseline model did well across horizons, while for Semi-durables, the media-based indices showed improvement in forecast performance at longer horizons. The MCS procedure also revealed how the median CRPS score could be misleading in determining forecasting performance if taken in isolation when many models are estimated. In these cases, the confidence of the error measure needs to be taken into account; not just a single observation. It was observed that for the median CRPS score, the Loughran-based sentiment indices were considered the best model for multiple sub-components of PCE, while in the MCS which incorporates block bootstrapping, the finding only held for total PCE.

Lastly, the performance of the different sentiment indices during the financial crisis were assessed. The Loughran-based models did well to improve the $t + 1$ forecasts of the sub-components of PCE within the period, but struggled to improve on the baseline for the overall changes in total PCE during the crisis.

The empirical results motivate for further research on how text-based analysis could be useful as official economic indicators in forecasting models.

6.3 Domain-specific dictionary generation using machine learning

In the fifth chapter of the thesis, against the backdrop of the success of text analysis illustrated empirically in chapter 3 and 4 using Loughran and McDonald (2011)'s lexicon, we demonstrated the benefit of applying domain-specific dictionaries when constructing sentiment indices. In chapter 5, machine learning was also evaluated as an alternative to the manual, labour-intensive process of constructing a domain-specific lexicon or dictionary. This was accomplished by operationalising subjective information processing using random forests.

In the procedure, tokens (words) were statistically selected using recursive feature elimination to generate a domain-specific dictionary from a corpus of text consisting of both uni-grams and bi-grams. The bi-grams themselves were investigated and shown to contain central ideas (topics) that aid in capturing domain-specific sentiment for the CCI, BCI, Civil Confidence index, and PMI[1] as well as the leading business cycle indicator constructed by the South African Reserve Bank (Leading). The central topics range from issues on government, business, and growth as well as what is happening in the stock markets. This concentration of tokens around key concepts could be linked to the hypothesis of Sims (2003)'s rational inattention model. Economic agents, having to focus on the economy and various intricacies surrounding it, face an opportunity cost connected to it. Agents would rather process information on economic activity (and central topics) through concentrated sources such as the media (and expert reporting), where key information on business, markets, government, etc. can be more easily digested. This distillation of key information can be tracked in a generalised construct known as sentiment indices. The empirical results showed that indices constructed from machine-generated dictionaries have a lower RMSE for a five-year holdout sample period, 2012 to 2017, compared to a sentiment index constructed from a commonly used financial dictionary.

The domain-specific generated indices significantly improved the fit over an index constructed using the Loughran dictionary. The largest improvement was

[1] These series are all survey-based sentiment indices constructed by the Bureau for Economic Research, Stellenbosch University

observed for the leading economic indicator (released by the SARB), where the domain-specific index improved the fit by a factor of 2.5. The results support the findings in Kalamara et al. (2020) where the combination of text based variables and machine learning techniques improved upon the baseline of an AR(1) model and standalone dictionary based methods when forecasting macroeconomic variables.

One of the other key findings of the chapter is that having a manually generated dictionary act as a prior for a domain narrows the tokens that the random forest has to search over, while maintaining the same lower RMSE out-of-sample as an unrestricted model that optimises over all tokens. Employing such a manually generated dictionary as a prior also decreases the computational burden in the pre-processing steps and consequently the estimation of the dictionaries. Practical considerations were also discussed when implementing the random forest algorithm tuning parameters. In the case of this thesis' data, the tuning parameters have no statistical difference in the RMSE across all different specifications, although this finding could be different for other studies. Future research should investigate the effects of tuning parameters on generating a dictionary when done in an automatic fashion, where tuning of the algorithms could drastically change the outcome. A better understanding of the dynamics between the tuning parameters and the resulting dictionary could improve the robustness of a dictionary, especially in an operational setting.

6.4 Conclusion

In conclusion, Romer (2016)'s controversial paper on the trouble with macroeconomics could not have come at a more interesting time. Economists are no longer standing idly by, solely focusing on deductive modelling (such as DSGE), but have embraced the changing landscape of applied research and the tools that have accompanied it, both within academia and industry.

This thesis explores text analysis as an example of one such area for the purposes of quantifying a measure of sentiment as projected by the media. It is argued in this thesis that it is feasible to construct a media-based sentiment index for South Africa that can replicate commonly used confidence measures. These newly constructed measures of sentiment offer a different perspective on consumer spending and savings behaviour that encapsulates news-driven business cycles. Besides capturing information on consumer sentiment, the media-based sentiment indices were shown to be effective in improving forecasts of personal consumption measures over the both short and long horizons. The study also questioned the commonly held belief that the Loughran dictionary is appropriate for capturing the nuance within economic reporting. By creating a domain-specific dictionary, the results show that applying a generalised financial dictionary such as Loughran introduces inefficiencies in information extraction.

The empirical findings in this dissertation have three important implications for both research and industry. Firstly, chapter 3 provides evidence that media-based sentiment indices could provide an complement and possible alternative to survey-based methods of consumer confidence. Secondly, 4 empirically showed that the incorporation of text-based sentiment into forecasting models not only contributes new information, but also overcomes certain practical constraints associated with survey-based sentiment indices. The inclusion of MSIs into various economic models shows promise to improve forecast results for consumption measures. Finally, applying machine learning algorithms to generate domain-specific sentiment dictionaries can greatly ease the challenge of extracting information from text. Evidence was presented that automatically generated dictionaries are shown to be more representative of underlying sentiment for multiple confidence indices tested. The algorithm presented in chapter 5 has the potential to be used for various economic or financial applications in a fully scalable production pipeline. These traits should especially be of interest to practitioners in industry that rely on advanced text analytics for forecasting.

These results contribute to the literature on the use of media text data as a source of economic information. These unconventional information sources may be used as an alternative (or complement) to already well-established economic indicators, or perhaps grow into a completely new set of economic indicators in their own right.

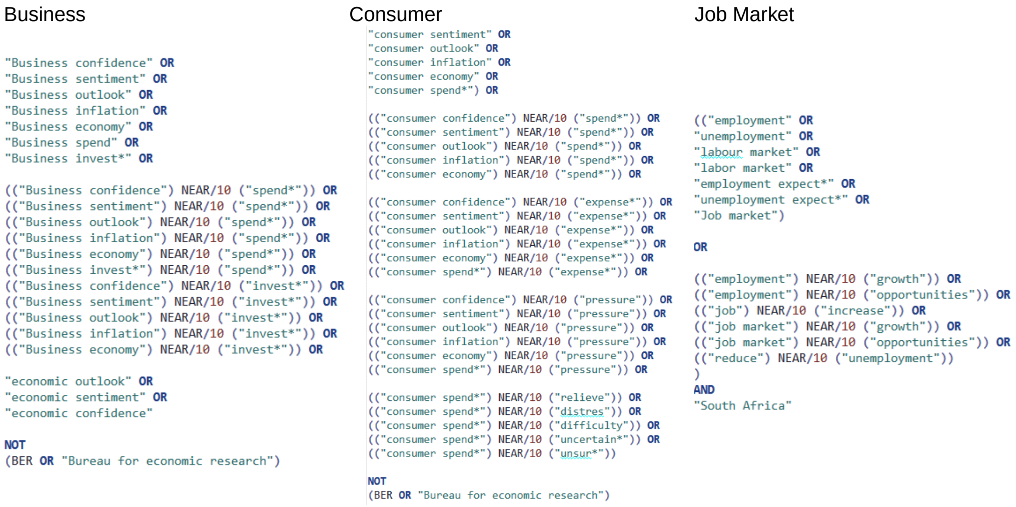
Appendices

Appendix A

Data overview

A.1 Meltwater boolean queries

To create the Meltwater indices, the author had to apply boolean searches to extract the articles of interest. Accordingly, three different queries were constructed to capture sentiment around: business, consumer and jobs.



Business

```
"Business confidence" OR
"Business sentiment" OR
"Business outlook" OR
"Business inflation" OR
"Business economy" OR
"Business spend" OR
"Business invest*" OR

(("Business confidence") NEAR/10 ("spend*")) OR
(("Business sentiment") NEAR/10 ("spend*")) OR
(("Business outlook") NEAR/10 ("spend*")) OR
(("Business inflation") NEAR/10 ("spend*")) OR
(("Business economy") NEAR/10 ("spend*")) OR
(("Business invest*") NEAR/10 ("spend*")) OR
(("Business confidence") NEAR/10 ("invest*")) OR
(("Business sentiment") NEAR/10 ("invest*")) OR
(("Business outlook") NEAR/10 ("invest*")) OR
(("Business inflation") NEAR/10 ("invest*")) OR
(("Business economy") NEAR/10 ("invest*")) OR

"economic outlook" OR
"economic sentiment" OR
"economic confidence"

NOT
(BER OR "Bureau for economic research")
```

Consumer

```
"consumer sentiment" OR
"consumer outlook" OR
"consumer inflation" OR
"consumer economy" OR
"consumer spend*") OR

(("consumer confidence") NEAR/10 ("spend*")) OR
(("consumer sentiment") NEAR/10 ("spend*")) OR
(("consumer outlook") NEAR/10 ("spend*")) OR
(("consumer inflation") NEAR/10 ("spend*")) OR
(("consumer economy") NEAR/10 ("spend*")) OR

(("consumer confidence") NEAR/10 ("expense*")) OR
(("consumer sentiment") NEAR/10 ("expense*")) OR
(("consumer outlook") NEAR/10 ("expense*")) OR
(("consumer inflation") NEAR/10 ("expense*")) OR
(("consumer economy") NEAR/10 ("expense*")) OR
(("consumer spend*") NEAR/10 ("expense*")) OR

(("consumer confidence") NEAR/10 ("pressure")) OR
(("consumer sentiment") NEAR/10 ("pressure")) OR
(("consumer outlook") NEAR/10 ("pressure")) OR
(("consumer inflation") NEAR/10 ("pressure")) OR
(("consumer economy") NEAR/10 ("pressure")) OR
(("consumer spend*") NEAR/10 ("pressure")) OR

(("consumer spend*") NEAR/10 ("relieve")) OR
(("consumer spend*") NEAR/10 ("distres")) OR
(("consumer spend*") NEAR/10 ("difficulty")) OR
(("consumer spend*") NEAR/10 ("uncertain*")) OR
(("consumer spend*") NEAR/10 ("unsur*"))

NOT
(BER OR "Bureau for economic research")
```

Job Market

```
(("employment" OR
"unemployment" OR
"labour market" OR
"labor market" OR
"employment expect*" OR
"unemployment expect*" OR
"Job market")

OR

(("employment") NEAR/10 ("growth")) OR
(("employment") NEAR/10 ("opportunities")) OR
(("job") NEAR/10 ("increase")) OR
(("job market") NEAR/10 ("growth")) OR
(("job market") NEAR/10 ("opportunities")) OR
(("reduce") NEAR/10 ("unemployment"))
)
AND
"South Africa"
```

Figure A.1: Boolean search criteria used to extract relevant articles

Appendix B

Test results

B.1 Unit root and cointegration tests

In chapter 3 a VAR analysis is conducted. For this test, the series in question need to be tested to see if they contain a unit root process. Given the small sample size, multiple unit root tests were conducted, namely ADF (Said and Dickey (1984)), Phillips-Perron (Phillips and Perron (1988)), and KPSS (Kwiatkowski et al. (1992))

For the Augmented Dickey-Fuller test (ADF), the general regression equation, which incorporates a constant and a linear trend, is used and the t-statistic for a first order autoregressive coefficient equals one is computed (Trapletti and Hornik 2019):

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{B.1}$$

The same specification is used for the Phillips–Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root test, which incorporates a constant and a linear trend.

Table B.1: Unit root test results

Variable	Statistic	P Value	Parameter	Method	Alternative
Cci	-3.00	0.18	3.00	Augmented Dickey-Fuller Test	stationary
Msi Per Cluster	-2.88	0.23	3.00	Augmented Dickey-Fuller Test	stationary
Msi Per Dtw	-1.64	0.71	3.00	Augmented Dickey-Fuller Test	stationary
News24	-2.25	0.47	3.00	Augmented Dickey-Fuller Test	stationary
Cci	0.09	0.10*	3.00	KPSS Test for Trend Stationarity	explosive
Msi Per Cluster	0.09	0.10*	3.00	KPSS Test for Trend Stationarity	explosive
Msi Per Dtw	0.13	0.07*	3.00	KPSS Test for Trend Stationarity	explosive
News24	0.09	0.10*	3.00	KPSS Test for Trend Stationarity	explosive
Cci	-21.86	0.98	3.00	Phillips-Perron Unit Root Test	explosive
Msi Per Cluster	-14.16	0.78	3.00	Phillips-Perron Unit Root Test	explosive
Msi Per Dtw	-10.87	0.56	3.00	Phillips-Perron Unit Root Test	explosive
News24	-17.97	0.94	3.00	Phillips-Perron Unit Root Test	explosive

Note: 0.001****, 0.01***, 0.05**, 0.1*

The series were also tested to see if they might be integrated. For this test we use the Johansen Test. Given a general VAR of the form:

$$X_t = \mu + \Phi D_t + \Pi_p X_{t-p} + \cdots + \Pi_1 X_{t-1} + e_t, \quad t = 1, \ldots, T \tag{B.2}$$

There are two possible specifications for error correction: either a long-run or transitory VECM. We perform an eigenvalue decomposition and inferences are drawn on Π using the cumulative long-run impacts.

Table B.2: Johanses cointegration test: MSI per cluster

hypothesis	10%	5%	1%	test
r <= 1	10.49	12.25	16.26	8.16
r = 0	16.85	18.96	23.65	20.19

Table B.3: Johanses cointegration test: MSI per DTW

hypothesis	10%	5%	1%	test
r <= 1	10.49	12.25	16.26	6.63
r = 0	16.85	18.96	23.65	9.53

Table B.4: Johanses cointegration test: MSI per News24

hypothesis	10%	5%	1%	test
r <= 1	10.49	12.25	16.26	9.44
r = 0	16.85	18.96	23.65	12.77

Tables B.2 - B.4 shows the maximal eigenvalue statistic for the two hypothesis of $r \leq 1$ and $r = 0$. In the case of testing whether the CCI and MSI per cluster is cointegrated, the test statistic exceeds the 1% significance level (table B.2). This would indicate that the matrix is of rank 1 and as such a linear combination of the two indices will result in a stationary series. This result however does not hold for MSI per dtw or MSI news24.

B.2 Toda-Yamamoto procedure

Toda and Yamamoto (1995)'s paper shows provides a 13-step procedure whereby you can estimate VARs, formulated in levels, and test general restrictions on the parameter matrices even if the processes may be integrated or cointegrated of an arbitrary order. The steps can be summarised as follow:[1]

1. Test each of the time-series to determine their order of integration. Ideally, this should involve using a test (such as the ADF test) for which the null hypothesis is non-stationarity; as well as a test (such as the KPSS test) for which the null is stationarity.

2. Let the maximum order of integration for the group of time-series be m. So, if there are two time-series and one is found to be I(1) and the other is I(2), then $m = 2$. If one is I(0) and the other is I(1), then $m = 1$, etc.

3. Set up a VAR model in the *levels* of the data, regardless of the orders of integration of the various time-series. Most importantly, you must not *difference* the data, no matter what you found at Step 1.

[1] An example and further discussion of these steps can be found in the original blog posted by Prof. David Giles, `https://davegiles.blogspot.com/2011/04/testing-for-granger-causality.html`

4. Determine the appropriate maximum lag length for the variables in the VAR, say p, using the usual methods. Specifically, base the choice of p on the usual information criteria, such as AIC, SIC.

5. Make sure that the VAR is well-specified and that there is no serial correlation in the residuals. If need be, increase p until any autocorrelation issues are resolved.

6. If two or more of the time-series have the same order of integration, at Step 1, then test to see if they are cointegrated, preferably using Johansen's methodology (based on your VAR) for a reliable result.

7. No matter what you conclude about cointegration at Step 6, this is not going to affect what follows. It just provides a possible cross-check on the validity of your results at the very end of the analysis.

8. Now take the preferred VAR model and add in m additional lags of each of the variables into each of the equations.

9. Test for Granger non-causality as follows: For expository purposes, suppose that the VAR has two equations, one for X and one for Y. Test the hypothesis that the coefficients of (only) the first p lagged values of X are zero in the Y equation, using a standard Wald test. Then do the same thing for the coefficients of the lagged values of Y in the X equation.

10. It's essential that you don't include the coefficients for the 'extra' m lags when you perform the Wald tests. They are there just to fix up the asymptotics.

11. The Wald test statistics will be asymptotically chi-square distributed with p d.o.f., under the null.

12. Rejection of the null implies a rejection of Granger non-causality. That is, a rejection supports the presence of Granger causality.

13. Finally, look back at what you concluded in Step 6 about cointegration

B.3 Model Confidence sets

Table B.5: Forecast ranking based on the median CRPS score for each of the model specifications.

Outcome	Models	Type	(t+1)	(t+2)	(t+3)	(t+4)	(t+6)	(t+8)	MeanRank
PCE Durable	Baseline	CRPS	7	3	3	4	2	**1**	3.33
PCE Durable	Baseline + CCI	CRPS	8	8	7	8	5	3	6.50
PCE Durable	Baseline + CCI + Henry	CRPS	6	5	2	2	4	7	4.33
PCE Durable	Baseline + CCI + Loughran	CRPS	4	7	8	6	8	8	6.83
PCE Durable	Baseline + CCI + PCA	CRPS	5	4	5	7	7	5	5.50
PCE Durable	Baseline + Henry	CRPS	**1**	2	**1**	**1**	**1**	4	1.67
PCE Durable	Baseline + Loughran	CRPS	2	6	6	3	6	6	4.83
PCE Durable	Baseline + PCA	CRPS	3	**1**	4	5	3	2	3.00
PCE Non Durable	Baseline	CRPS	8	2	5	**1**	5	4	4.17
PCE Non Durable	Baseline + CCI	CRPS	7	3	4	6	7	7	5.67
PCE Non Durable	Baseline + CCI + Henry	CRPS	5	5	7	8	8	8	6.83
PCE Non Durable	Baseline + CCI + Loughran	CRPS	6	7	2	2	2	2	3.50
PCE Non Durable	Baseline + CCI + PCA	CRPS	2	**1**	3	7	6	6	4.17
PCE Non Durable	Baseline + Henry	CRPS	3	6	8	5	4	5	5.17
PCE Non Durable	Baseline + Loughran	CRPS	**1**	8	**1**	3	**1**	**1**	2.50
PCE Non Durable	Baseline + PCA	CRPS	4	4	6	4	3	3	4.00
PCE Semi Durable	Baseline	CRPS	4	6	5	7	4	**1**	4.50
PCE Semi Durable	Baseline + CCI	CRPS	6	**1**	3	6	3	4	3.83
PCE Semi Durable	Baseline + CCI + Henry	CRPS	**1**	2	4	5	8	8	4.67
PCE Semi Durable	Baseline + CCI + Loughran	CRPS	7	5	**1**	**1**	**1**	6	3.50
PCE Semi Durable	Baseline + CCI + PCA	CRPS	5	4	2	2	5	7	4.17
PCE Semi Durable	Baseline + Henry	CRPS	2	7	8	8	7	5	6.17
PCE Semi Durable	Baseline + Loughran	CRPS	8	3	7	3	2	2	4.17
PCE Semi Durable	Baseline + PCA	CRPS	3	8	6	4	6	3	5.00
PCE Services	Baseline	CRPS	7	4	8	7	2	**1**	4.83
PCE Services	Baseline + CCI	CRPS	3	7	**1**	**1**	3	3	3.00
PCE Services	Baseline + CCI + Henry	CRPS	5	5	3	3	4	6	4.33
PCE Services	Baseline + CCI + Loughran	CRPS	8	6	2	2	**1**	7	4.33
PCE Services	Baseline + CCI + PCA	CRPS	6	8	4	4	5	5	5.33
PCE Services	Baseline + Henry	CRPS	**1**	3	7	6	8	2	4.50
PCE Services	Baseline + Loughran	CRPS	4	2	5	8	7	8	5.67
PCE Services	Baseline + PCA	CRPS	2	**1**	6	5	6	4	4.00
PCE Tot	Baseline	CRPS	2	8	6	3	3	**1**	3.83
PCE Tot	Baseline + CCI	CRPS	8	5	3	6	5	2	4.83
PCE Tot	Baseline + CCI + Henry	CRPS	**1**	7	4	4	6	5	4.50
PCE Tot	Baseline + CCI + Loughran	CRPS	4	4	**1**	**1**	2	8	3.33
PCE Tot	Baseline + CCI + PCA	CRPS	7	3	5	7	8	7	6.17
PCE Tot	Baseline + Henry	CRPS	6	6	7	5	4	3	5.17
PCE Tot	Baseline + Loughran	CRPS	5	**1**	2	2	**1**	6	2.83
PCE Tot	Baseline + PCA	CRPS	3	2	8	8	7	4	5.33

Table B.6 and 4.3 display the results from the MCS procedure that tests the hypothesis of EPA. For total personal consumption expenditure, the SSM consists of single model: baseline + Loughran. A single model SSM is however not the case when we examine the sub-components of PCE. Durables, for instance, contain the Loughran and Baseline model for its $t + 1$ forecast with a high p-value of 0.99.

This means that when forecasting Durables, it is very difficult to improve on the baseline model which consists of the outcome variable and disposable income. The results for Durables do not contain any of the CCI specifications. In contrast to the median CRPS scores, the SSM for Services shows that the traditional CCI measure does very well in the forecast horizon $h = 1$. This result is in line with the result of Gelper, Lemmens, and Croux (2007), who found that the CCI was well suited to the forecasting of services. The p-values for horizons $h = \{2, 3\}$ rejects the null hypothesis of equal predictive ability, suggesting that there is another model that could perform just as well or better.

In the remaining SSM, the results include the baseline and at least one of the sentiment indices. For Non-durables, the short-term forecast, $t = 1$, is very difficult to forecast and the SSM contains almost all the models. In the longer horisons, $t = \{2, 3, 4\}$, the media-based indices perform better. In terms of Semi-durables, the SSM considers a lot of models that contain the traditional CCI, all having equal predictive power.

Table B.6: Model confidence set results.

Outcome	H	Model	Rank R	V R	MCS R	Loss	P Value
PCE Durable	1	Baseline + Loughran	1.00	-0.01	1.00	0.34	0.99
PCE Durable	1	Baseline	2.00	0.01	0.99	0.34	0.99
PCE Durable	2	Baseline	1.00	-0.35	1.00	0.41	0.73
PCE Durable	3	Baseline	1.00	-0.03	1.00	0.45	0.98
PCE Durable	3	Baseline + Loughran	2.00	0.03	0.98	0.45	0.98
PCE Durable	4	Baseline	1.00	-1.13	1.00	0.48	0.26
PCE Durable	6	Baseline	1.00	-0.86	1.00	0.49	0.42
PCE Durable	8	Baseline	1.00	-0.86	1.00	0.45	0.42
PCE Non Durable	1	Baseline + Loughran	1.00	-0.21	1.00	0.45	0.90
PCE Non Durable	1	Baseline	2.00	0.21	1.00	0.46	0.90
PCE Non Durable	1	Baseline + PCA	3.00	0.28	1.00	0.46	0.90
PCE Non Durable	1	Baseline + CCI	4.00	0.30	1.00	0.46	0.90
PCE Non Durable	1	Baseline + CCI + PCA	5.00	0.58	1.00	0.46	0.90
PCE Non Durable	1	Baseline + CCI + Loughran	6.00	0.82	0.90	0.46	0.90
PCE Non Durable	2	Baseline + Loughran	1.00	-0.04	1.00	0.49	0.96
PCE Non Durable	2	Baseline + PCA	2.00	0.04	0.96	0.49	0.96
PCE Non Durable	3	Baseline + Loughran	1.00	-0.21	1.00	0.54	0.83
PCE Non Durable	4	Baseline	1.00	-0.06	1.00	0.57	0.96
PCE Non Durable	4	Baseline + Loughran	2.00	0.06	0.96	0.57	0.96
PCE Non Durable	6	Baseline + Loughran	1.00	-0.78	1.00	0.56	0.44
PCE Non Durable	8	Baseline	1.00	-0.31	1.00	0.52	0.77
PCE Semi Durable	1	Baseline	1.00	-0.44	1.00	0.52	0.93
PCE Semi Durable	1	Baseline + Loughran	2.00	0.44	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + CCI	3.00	0.56	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + PCA	4.00	0.59	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + Henry	5.00	0.78	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + CCI + Loughran	6.00	0.79	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + CCI + PCA	7.00	0.79	1.00	0.53	0.93
PCE Semi Durable	1	Baseline + CCI + Henry	8.00	1.25	0.74	0.53	0.93
PCE Semi Durable	2	Baseline + Loughran	1.00	-0.15	1.00	0.59	0.95
PCE Semi Durable	2	Baseline + CCI + Loughran	2.00	0.15	1.00	0.59	0.95
PCE Semi Durable	2	Baseline + CCI + PCA	3.00	0.33	1.00	0.59	0.95
PCE Semi Durable	2	Baseline + PCA	4.00	0.36	1.00	0.59	0.95
PCE Semi Durable	2	Baseline + CCI	5.00	0.45	1.00	0.59	0.95
PCE Semi Durable	2	Baseline	6.00	0.57	0.99	0.59	0.95
PCE Semi Durable	3	Baseline + CCI + Loughran	1.00	-0.14	1.00	0.56	0.89
PCE Semi Durable	4	Baseline + CCI + Loughran	1.00	-0.18	1.00	0.60	0.86
PCE Semi Durable	6	Baseline + CCI + Loughran	1.00	-1.26	1.00	0.59	0.21
PCE Semi Durable	8	Baseline + CCI	1.00	-0.19	1.00	0.64	0.86
PCE Services	1	Baseline + CCI	1.00	-0.74	1.00	0.39	0.47
PCE Services	2	Baseline + CCI	1.00	-3.00	1.00	0.42	0.00
PCE Services	3	Baseline + CCI	1.00	-3.13	1.00	0.43	0.00
PCE Services	4	Baseline + CCI	1.00	-0.08	1.00	0.43	0.96
PCE Services	4	Baseline	2.00	0.08	1.00	0.43	0.96
PCE Services	4	Baseline + Loughran	3.00	0.34	1.00	0.43	0.96
PCE Services	4	Baseline + PCA	4.00	0.37	0.99	0.43	0.96
PCE Services	6	Baseline + PCA	1.00	-0.03	1.00	0.41	0.98
PCE Services	6	Baseline + Loughran	2.00	0.03	0.98	0.41	0.98
PCE Services	8	Baseline	1.00	-1.57	1.00	0.43	0.12
PCE Tot	1	Baseline + Loughran	1.00	-1.44	1.00	0.40	0.15
PCE Tot	2	Baseline + Loughran	1.00	-1.09	1.00	0.49	0.30
PCE Tot	3	Baseline + Loughran	1.00	-1.06	1.00	0.55	0.33
PCE Tot	4	Baseline + Loughran	1.00	-1.15	1.00	0.62	0.26
PCE Tot	6	Baseline + Loughran	1.00	-0.96	1.00	0.63	0.40
PCE Tot	8	Baseline + Loughran	1.00	-0.68	1.00	0.61	0.51

Appendix C

Dictionary overview

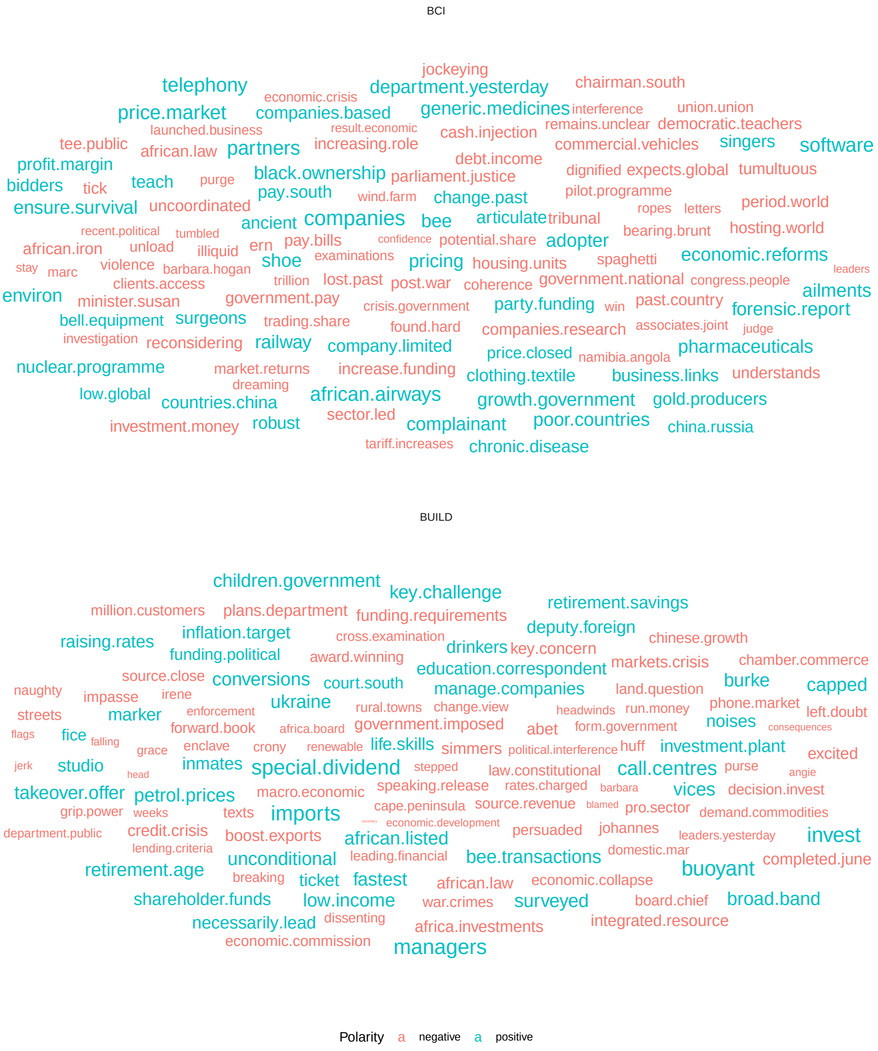


Figure C.1: Sample of 150 words from the different dictionaries created. Polarity of word was calculated as mean score over all model specifications.

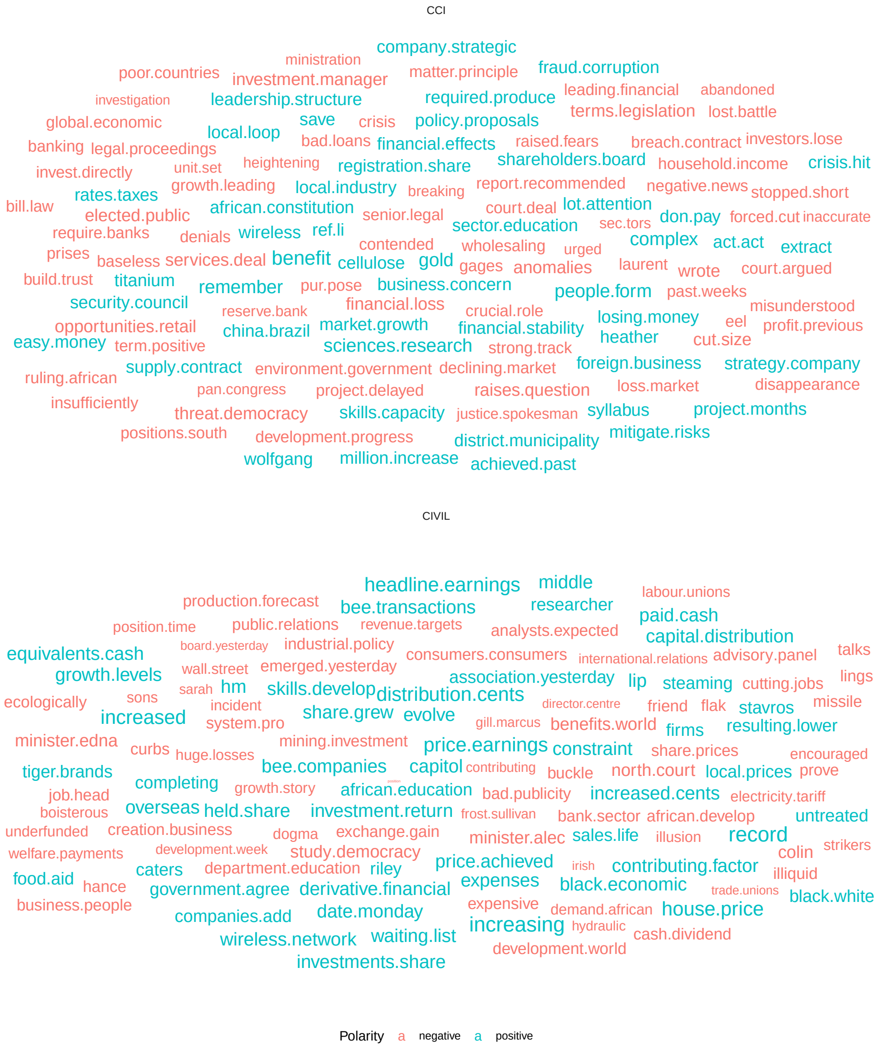


Figure C.2: Sample of 150 words from the different dictionaries created. Polarity of word was calculated as mean score over all model specifications.

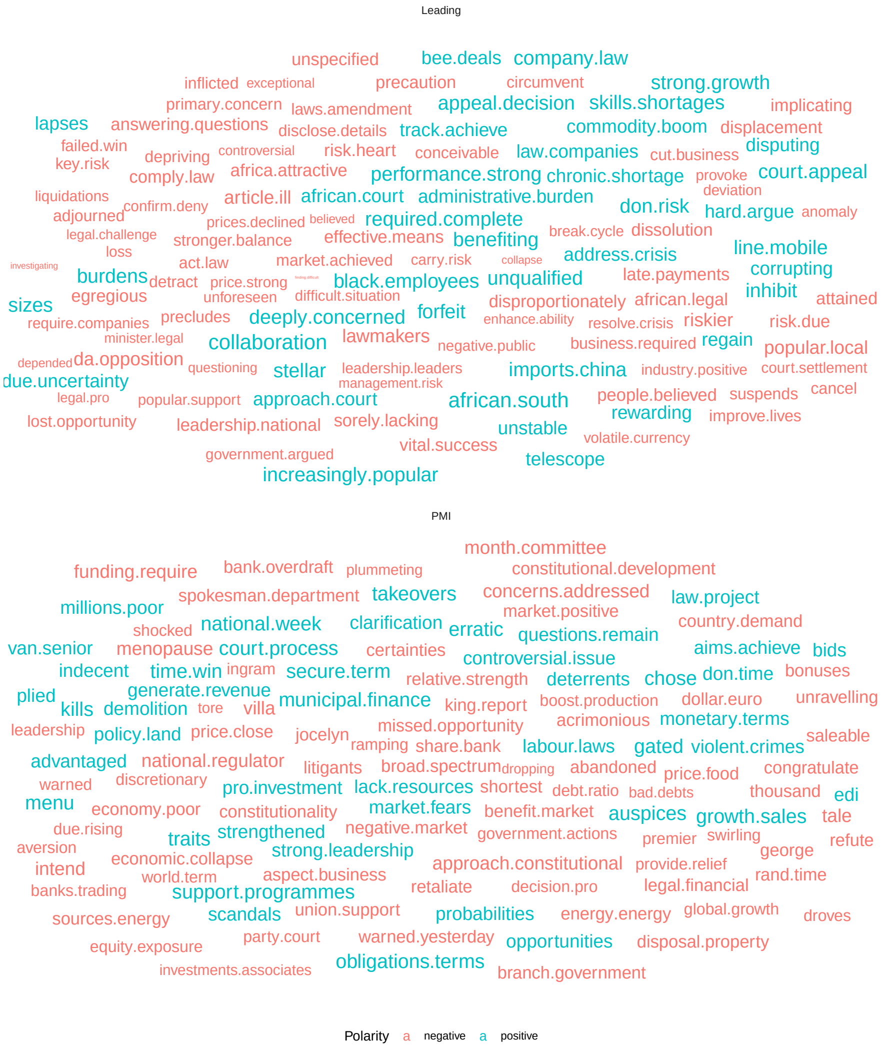


Figure C.3: Sample of 150 words from the different dictionaries created. Polarity of word was calculated as mean score over all model specifications.

Appendix D

Software information

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
## LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3
##
## locale:
##  [1] LC_CTYPE=en_BW.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_BW.UTF-8        LC_COLLATE=en_BW.UTF-8
##  [5] LC_MONETARY=en_BW.UTF-8    LC_MESSAGES=en_BW.UTF-8
##  [7] LC_PAPER=en_BW.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_BW.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
```

```
##  [1] ranger_0.11.2           corrr_0.4.2             doMC_1.3.6
##  [4] doParallel_1.0.15       iterators_1.0.12        foreach_1.4.7
##  [7] quanteda_1.5.2          SnowballC_0.6.0         SentimentAnalysis_1.3-3
## [10] xts_0.12-0              dlm_1.1-5               png_0.1-7
## [13] zoo_1.8-7               xtable_1.8-4            TSclust_1.2.4
## [16] cluster_2.1.0           pdc_1.0.3               wmtsa_2.0-3
## [19] stargazer_5.2.2         scales_1.1.0            snakecase_0.11.0
## [22] NewsR_0.1.0             knitr_1.26              patchwork_1.0.0
## [25] gridExtra_2.3           ggpubr_0.2.4            magrittr_1.5
## [28] glue_1.4.1              ggraph_2.0.0            igraph_1.2.4.2
## [31] ggthemes_4.2.0          ggsci_2.9               gghighlight_0.1.0
## [34] ggfortify_0.4.8         forecast_8.10           FactoMineR_2.0
## [37] factoextra_1.0.6        corrplot_0.84           aws.s3_0.3.12
## [40] openxlsx_4.1.4          tbl2xts_0.1.3           lubridate_1.7.4
## [43] forcats_0.5.0           stringr_1.4.0           dplyr_0.8.5
## [46] purrr_0.3.4             readr_1.3.1             tidyr_1.0.0
## [49] tibble_3.0.3            ggplot2_3.3.1           tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##   [1] readxl_1.3.1            backports_1.1.5         fastmatch_1.1-0
##   [4] lazyeval_0.2.2          crosstalk_1.0.0         splus2R_1.2-2
##   [7] digest_0.6.25           htmltools_0.4.0         viridis_0.5.1
##  [10] fansi_0.4.1             aod_1.3.1               aws.signature_0.5.2
##  [13] graphlayouts_0.5.0      modelr_0.1.5            RcppParallel_4.4.4
##  [16] sandwich_2.5-1          tseries_0.10-47         strucchange_1.5-2
##  [19] colorspace_1.4-1        rvest_0.3.5             ggrepel_0.8.1
##  [22] haven_2.2.0             xfun_0.11               crayon_1.3.4
##  [25] jsonlite_1.6.1          polyclip_1.10-0         stopwords_1.0
##  [28] gtable_0.3.0            webshot_0.5.2           spacyr_1.2.1
##  [31] quantmod_0.4-15         DBI_1.1.0               miniUI_0.1.1.1
##  [34] Rcpp_1.0.5              dtw_1.21-3              viridisLite_0.3.0
##  [37] flashClust_1.01-2       proxy_0.4-23            longitudinalData_2.4.1
##  [40] htmlwidgets_1.5.1       httr_1.4.1              ggwordcloud_0.5.0
##  [43] ellipsis_0.3.1          pkgconfig_2.0.3         farver_2.0.1
##  [46] nnet_7.3-14             dbplyr_1.4.2            labeling_0.3
##  [49] tidyselect_1.1.0        rlang_0.4.7             manipulateWidget_0.10.0
##  [52] later_1.0.0             munsell_0.5.0           cellranger_1.1.0
##  [55] tools_3.6.3             cli_2.0.2               generics_0.0.2
##  [58] broom_0.5.2             evaluate_0.14           locpol_0.7-0
##  [61] fastmap_1.0.1           yaml_2.2.0              fs_1.3.1
##  [64] tidygraph_1.1.2         zip_2.0.4               rgl_0.100.30
##  [67] dendextend_1.13.2       clv_0.3-2.1             nlme_3.1-147
##  [70] mime_0.9                leaps_3.0               xml2_1.2.2
##  [73] compiler_3.6.3          rstudioapi_0.10         curl_4.3
##  [76] ggsignif_0.6.0          reprex_0.3.0            tweenr_1.0.1
##  [79] stringi_1.4.6           lattice_0.20-41         Matrix_1.2-18
##  [82] urca_1.3-0              vctrs_0.3.1             vars_1.5-3
##  [85] pillar_1.4.6            lifecycle_0.2.0         lmtest_0.9-37
##  [88] data.table_1.12.8       cowplot_1.0.0           httpuv_1.5.2
##  [91] R6_2.4.1                promises_1.1.0          KernSmooth_2.23-17
##  [94] codetools_0.2-16        MASS_7.3-51.6           assertthat_0.2.1
##  [97] withr_2.1.2             fracdiff_1.5-0          hms_0.5.2
## [100] quadprog_1.5-8          timeDate_3043.102       class_7.3-17
## [103] rmarkdown_2.3           misc3d_0.8-4            ifultools_2.0-5
## [106] TTR_0.23-6              ggforce_0.3.1           scatterplot3d_0.3-41
## [109] shiny_1.4.0             base64enc_0.1-3
```

References

Abberger, Klaus. 2006. “Another Look at the Ifo Business Cycle Clock.” Journal of Business Cycle Measurement and Analysis 2005 (3): 431–43.

Acemoglu, Daron, and Andrew Scott. 1994. “Consumer Confidence and Rational Expectations: Are Agents’ Beliefs Consistent with the Theory?” The Economic Journal, 1–19.

Aggarwal, Charu C, and ChengXiang Zhai. 2012. Mining Text Data. Springer Science & Business Media.

Ahmed, M Iqbal, and Steven P Cassou. 2016. “Does Consumer Confidence Affect Durable Goods Spending During Bad and Good Economic Times Equally?” Journal of Macroeconomics 50: 86–97.

Akerlof, George A, William T Dickens, George L Perry, Truman F Bewley, and Alan S Blinder. 2000. “Near-Rational Wage and Price Setting and the Long-Run Phillips Curve.” Brookings Papers on Economic Activity 2000 (1): 1–60.

Angeletos, George-Marios, and Jennifer La’O. 2013. “Sentiments.” Econometrica 81 (2): 739–79.

Antweiler, Werner, and Murray Z Frank. 2004. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” The Journal of Finance 59 (3): 1259–94.

Ardia, David, Keven Bluteau, and Kris Boudt. 2019. “Questioning the News About Economic Growth: Sparse Forecasting Using Thousands of News-Based Sentiment Values.” International Journal of Forecasting.

Athey, Susan. 2015. “Machine Learning and Causal Inference for Policy Evaluation.” In Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 5–6. ACM.

———. 2017. "Beyond Prediction: Using Big Data for Policy Problems." Science 355 (6324): 483–85.

———. 2018. "The Impact of Machine Learning on Economics." In The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu. 2019. "Ensemble Methods for Causal Effects in Panel Data Settings." National Bureau of Economic Research.

Athey, Susan, and Guido Imbens. 2019. "Machine Learning Methods Economists Should Know About." arXiv Preprint arXiv:1903.10075.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2016. "Generalized Random Forests." `http://arxiv.org/abs/1610.01271`.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring Economic Policy Uncertainty." The Quarterly Journal of Economics 131 (4): 1593–1636.

Barsky, Robert B, and Eric R Sims. 2011. "News Shocks and Business Cycles." Journal of Monetary Economics 58 (3): 273–89.

———. 2012. "Information, Animal Spirits, and the Meaning of Innovations in Consumer Confidence." American Economic Review 102 (4): 1343–77.

Beaudry, Paul, and Franck Portier. 2004. "An Exploration into Pigou's Theory of Cycles." Journal of Monetary Economics 51 (6): 1183–1216.

———. 2014. "News-Driven Business Cycles: Insights and Challenges." Journal of Economic Literature 52 (4): 993–1074.

Benhabib, Jess, Pengfei Wang, and Yi Wen. 2015. "Sentiments and Aggregate Demand Fluctuations." Econometrica 83 (2): 549–85.

Bernanke, Ben. 2004. "Fedspeak." In Remarks at the Meetings of the American Economic Association, San Diego, Available at: Http://Www. Federalreserve. Gov/Boarddocs/Speeches/2004/200401032/Default. Htm.

Bernardi, Mauro, and Leopoldo Catania. 2018. "The Model Confidence Set Package for R." International Journal of Computational Economics and Econometrics 8 (2): 144–58.

Berndt, Donald J, and James Clifford. 1994. "Using Dynamic Time Warping to Find Patterns in Time Series." In KDD Workshop, 10:359–70. 16. Seattle, WA.

Biau, Olivier, and Angela D'Elia. 2009. "Euro Area Gdp Forecasting Using Large Survey Datasets." In A Random Forest Approach. Availabile via: Http://Unstats. Un. Org/Unsd/Nationalaccount/Workshops/2010/Moscow/Ac223-S73bk4. PDF.

Binge, Laurie H. 2018. "Methods for Aggregating Microeconomic Data: Applications to Art Prices, Business Sentiment and Historical Commodity Prices." PhD thesis, Stellenbosch: Stellenbosch University.

Blanchard, Olivier. 1993. "Consumption and the Recession of 1990-1991." The American Economic Review 83 (2): 270–74.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." Journal of Computational Science 2 (1): 1–8.

Bollerslev, Tim, Robert F Engle, and Daniel B Nelson. 1994. "ARCH Models." Handbook of Econometrics 4: 2959–3038.

Born, Benjamin, Michael Ehrmann, and Marcel Fratzscher. 2014. "Central Bank Communication on Financial Stability." The Economic Journal 124 (577): 701–34.

Bouchet-Valat, Milan. 2019. SnowballC: Snowball Stemmers Based on the c 'Libstemmer' Utf-8 Library. https://CRAN.R-project.org/package=SnowballC.

Bram, Jason, and Sydney Ludvigson. 1997. "Does Consumer Confidence Forecast Household Expenditure? A Sentiment Index Horse Race."

Breiman, Leo. 2001a. "Random Forests." Machine Learning 45 (1): 5–32.

———. 2001b. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." Statistical Science 16 (3): 199–231.

Carriero, Andrea, Todd E Clark, and Massimiliano Marcellino. 2015. "Bayesian Vars: Specification Choices and Forecast Accuracy." Journal of Applied Econometrics 30 (1): 46–73.

Carroll, Christopher D. 2003. "Macroeconomic Expectations of Households and Professional Forecasters." The Quarterly Journal of Economics 118 (1): 269–98.

Carroll, Christopher D, Jeffrey C Fuhrer, and David W Wilcox. 1994. "Does Consumer Sentiment Forecast Household Spending? If so, Why?" The American Economic Review 84 (5): 1397–1408.

Cavallo, Alberto. 2013. "Online and Official Price Indexes: Measuring Argentina's Inflation." Journal of Monetary Economics 60 (2): 152–65.

Chai, Tianfeng, and Roland R Draxler. 2014. "Root Mean Square Error (Rmse) or Mean Absolute Error (Mae)?–Arguments Against Avoiding Rmse in the Literature." Geoscientific Model Development 7 (3): 1247–50.

Chakraborty, Chiranjit, and Andreas Joseph. 2017. "Machine Learning at Central Banks."

Chen, Danqi, and Christopher Manning. 2014. "A Fast and Accurate Dependency Parser Using Neural Networks." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (Emnlp), 740–50.

Clark, Todd E, and Michael W McCracken. 2008. "Chapter 3 Forecasting with Small Macroeconomic Vars in the Presence of Instabilities." In Forecasting in the Presence of Structural Breaks and Model Uncertainty, 93–147. Emerald Group Publishing Limited.

Coertjens, Liesje, Vincent Donche, Sven De Maeyer, Gert Vanthournout, and Peter Van Petegem. 2012. "Longitudinal Measurement Invariance of Likert-Type Learning Strategy Scales: Are We Using the Same Ruler at Each Wave?" Journal of Psychoeducational Assessment 30 (6): 577–87.

Commission, Joint Research Centre-European. 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD publishing.

Corduas, Marcella. 2010. "Mining Time Series Data: A Selective Survey." In Data Analysis and Classification, 355–62. Springer.

Cowles, Alfred. 1933. "Can Stock Market Forecasters Forecast?" Econometrica: Journal of the Econometric Society, 309–24.

Curtin, Richard. 2007. "Consumer Sentiment Surveys: Worldwide Review and Assessment." OECD Journal. Journal of Business Cycle Measurement and Analysis 2007 (1): 7.

Daas, Piet JH, and Marco JH Puts. 2014. "Social Media Sentiment and Consumer Confidence." ECB Statistics Paper.

Del Negro, Marco, and Frank Schorfheide. 2004. "Priors from General Equilibrium Models for Vars." International Economic Review 45 (2): 643–73.

Diebold, Francis X. 2015. "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests." Journal of Business & Economic Statistics 33 (1): 1–1.

Diebold, Francis X, and Jose A Lopez. 1996. “Forecast Evaluation and Combination.” Handbook of Statistics 14: 241–68.

Diebold, Francis X, and Robert S Mariano. 2002. “Comparing Predictive Accuracy.” Journal of Business & Economic Statistics 20 (1): 134–44.

Dominitz, Jeff, and Charles F Manski. 2004. “How Should We Measure Consumer Confidence?” Journal of Economic Perspectives 18 (2): 51–66.

Doms, Mark E, and Norman J Morin. 2004. “Consumer Sentiment, the Economy, and the News Media.”

D’Orazio, Paola. 2017. “Big Data and Complexity: Is Macroeconomics Heading Toward a New Paradigm?” Journal of Economic Methodology 24 (4): 410–29.

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A Parsons. 2012. “Journalists and the Stock Market.” The Review of Financial Studies 25 (3): 639–79.

ECB. 2013. Confidence Indicators and Economic Developments. Vol. January. ECB Monthly Bulletin.

Economic Co-operation, Organisation for, and Development. 2003. Business Tendency Surveys: A Handbook. OECD Publishing.

Fan, Chengze Simon, and Phoebe Wong. 1998. “Does Consumer Sentiment Forecast Household Spending?: The Hong Kong Case.” Economics Letters 58 (1): 77–84.

Fraiberger, Samuel. 2016. “News Sentiment and Cross-Country Fluctuations.”

Freeman, Linton C. 1977. “A Set of Measures of Centrality Based on Betweenness.” Sociometry, 35–41.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. The Elements of Statistical Learning. Vol. 1. Springer series in statistics New York.

Friedman, Milton. 1957. “The Permanent Income Hypothesis.” In A Theory of the Consumption Function, 20–37. Princeton University Press.

Fuhrer, Jeffrey C. 1993. “What Role Does Consumer Sentiment Play in the Us Macroeconomy?” New England Economic Review, no. Jan: 32–44.

Gelper, Sarah, and Christophe Croux. 2010. “On the Construction of the European Economic Sentiment Indicator.” Oxford Bulletin of Economics and Statistics 72 (1): 47–62.

Gelper, Sarah, AurÃľlie Lemmens, and Christophe Croux. 2007. “Consumer Sentiment and Consumer Spending: Decomposing the Granger Causal Relationship in the Time Domain.” Applied Economics 39 (1): 1–11.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. “Text as Data.” National Bureau of Economic Research.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. “Variable Selection Using Random Forests.” Pattern Recognition Letters 31 (14): 2225–36.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics Using Search Engine Query Data.” Nature 457 (7232): 1012.

Giorgino, Toni. 2009. “Computing and Visualizing Dynamic Time Warping Alignments in R: The Dtw Package.” Journal of Statistical Software 31 (7): 1–24.

Girardi, Alessandro, and Andreas Reuter. 2016. “New Uncertainty Measures for the Euro Area Using Survey Data.” Oxford Economic Papers 69 (1): 278–300.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E Raftery. 2007. “Probabilistic Forecasts, Calibration and Sharpness.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2): 243–68.

Gneiting, Tilmann, and Adrian E Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” Journal of the American Statistical Association 102 (477): 359–78.

Gneiting, Tilmann, and Roopesh Ranjan. 2011. “Comparing Density Forecasts Using Threshold-and Quantile-Weighted Scoring Rules.” Journal of Business & Economic Statistics 29 (3): 411–22.

Goidel, Robert K, and Ronald E Langley. 1995. “Media Coverage of the Economy and Aggregate Economic Evaluations: Uncovering Evidence of Indirect Media Effects.” Political Research Quarterly 48 (2): 313–28.

Goldberg, Yoav. 2016. “A Primer on Neural Network Models for Natural Language Processing.” Journal of Artificial Intelligence Research 57: 345–420.

GrÃűmping, Ulrike. 2009. “Variable Importance Assessment in Regression: Linear Regression Versus Random Forest.” The American Statistician 63 (4): 308–19.

Guyon, Isabelle, and AndrÃĺ Elisseeff. 2003. "An Introduction to Variable and Feature Selection." Journal of Machine Learning Research 3 (Mar): 1157–82.

Hall, Robert E. 1978. "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence." Journal of Political Economy 86 (6): 971–87.

Hansen, Peter R. 2005. "A Test for Superior Predictive Ability." Journal of Business & Economic Statistics 23 (4): 365–80.

Hansen, Peter R, and Asger Lunde. 2005. "A Forecast Comparison of Volatility Models: Does Anything Beat a Garch (1, 1)?" Journal of Applied Econometrics 20 (7): 873–89.

Hansen, Peter R, Asger Lunde, and James M Nason. 2003. "Choosing the Best Volatility Models: The Model Confidence Set Approach." Oxford Bulletin of Economics and Statistics 65: 839–61.

———. 2011. "The Model Confidence Set." Econometrica 79 (2): 453–97.

Hansen, Stephen, and Michael McMahon. 2016. "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication." Journal of International Economics 99: S114–S133.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2014. "Transparency and Deliberation Within the Fomc: A Computational Linguistics Approach." The Quarterly Journal of Economics.

Harris, Zellig S. 1954. "Distributional Structure." Word 10 (2-3): 146–62.

Harvey, Andrew, and Chia-Hui Chung. 2000. "Estimating the Underlying Change in Unemployment in the Uk." Journal of the Royal Statistical Society: Series A (Statistics in Society) 163 (3): 303–9.

He, Zonglu, and Koichi Maekawa. 2001. "On Spurious Granger Causality." Economics Letters 73 (3): 307–13.

Henry, Elaine. 2008. "Are Investors Influenced by How Earnings Press Releases Are Written?" The Journal of Business Communication (1973) 45 (4): 363–407.

Hu, Minqing, and Bing Liu. 2004. "Mining and Summarizing Customer Reviews." In Proceedings of the Tenth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 168–77. ACM.

Hubert, Lawrence J, and Joel R Levin. 1976. "A General Statistical Framework for Assessing Categorical Clustering in Free Recall." Psychological Bulletin 83 (6):

1072.

Jacobs, Gilles, Els Lefever, and Veronique Hoste. 2018. "Economic Event Detection in Company-Specific News Text." In Proceedings of the First Workshop on Economics and Natural Language Processing, 1–10. Melbourne, Australia: Association for Computational Linguistics. `https://www.aclweb.org/anthology/W18-3101`.

Jegadeesh, Narasimhan, and Di Wu. 2013. "Word Power: A New Approach for Content Analysis." Journal of Financial Economics 110 (3): 712–29.

Johansen, SÃÿren. 1991. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models." Econometrica: Journal of the Econometric Society, 1551–80.

Jordan, Alexander, Fabian Krueger, and Sebastian Lerch. 2017. ScoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts. `https://CRAN.R-project.org/package=scoringRules`.

Kabundi, Alain, Elmarie Nel, and Franz Ruch. 2016. "Nowcasting Real Gdp Growth in South Africa." Economic Research Southern Africa.[Working Paper, No. 581.] Pretoria: National Treasury of South Africa.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia. 2020. "Making Text Count: Economic Forecasting Using Newspaper Text." Bank of England.

Kane, Michael J, Natalie Price, Matthew Scotch, and Peter Rabinowitz. 2014. "Comparison of Arima and Random Forest Time Series Models for Prediction of Avian Influenza H5n1 Outbreaks." BMC Bioinformatics 15 (1): 276.

Kershoff, George. 2000. "Measuring Business and Consumer Confidence in South Africa." BER, Stellenbosh, December.

Keynes, John Maynard. 1937. "The General Theory of Employment." The Quarterly Journal of Economics 51 (2): 209–23.

Khumalo, John. 2014. "Consumer Spending and Consumer Confidence in South Africa: Cointegration Analysis." Journal of Economics and Behavioral Studies 6 (2): 95.

Koop, Gary M. 2013. "Forecasting with Medium and Large Bayesian Vars." Journal of Applied Econometrics 28 (2): 177–203.

Koopman, S. J., and J. Durbin. 2003. "Filtering and smoothing of state vector for diffuse state-space models." Journal of Time Series Analysis 24 (1): 85–98. `http://ideas.repec.org/a/bla/jtsera/v24y2003i1p85-98.html`.

Krueger, Fabian, Sebastian Lerch, Thordis L Thorarinsdottir, and Tilmann Gneiting. 2016. "Probabilistic Forecasting and Comparative Model Assessment Based on Markov Chain Monte Carlo Output." arXiv Preprint arXiv:1608.06802.

Kuhn, Max, and Kjell Johnson. 2013. Applied Predictive Modeling. Vol. 26. Springer.

Kwiatkowski, Denis, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. 1992. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" Journal of Econometrics 54 (1-3): 159–78.

Labille, Kevin, Susan Gauch, and Sultan Alfarhood. 2017. "Creating Domain-Specific Sentiment Lexicons via Text Mining." In Proc. Workshop Issues Sentiment Discovery Opinion Mining (Wisdom).

Laio, Francesco, and Stefania Tamea. 2007. "Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables." Hydrology and Earth System Sciences Discussions 11 (4): 1267–77.

Lamla, Michael J, and Sarah M Lein. 2014. "The Role of Media for Consumers' Inflation Expectation Formation." Journal of Economic Behavior & Organization 106: 62–77.

Larsen, Vegard H, and Leif Anders Thorsrud. 2015. "The Value of News."

Leduc, Sylvain, Keith Sill, and Tom Stark. 2007. "Self-Fulfilling Expectations and the Inflation of the 1970s: Evidence from the Livingston Survey." Journal of Monetary Economics 54 (2): 433–59.

Lin, Jessica, and Yuan Li. 2009. "Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation." In International Conference on Scientific and Statistical Database Management, 461–77. Springer.

Liu, Bing. 2012. "Sentiment Analysis and Opinion Mining." Synthesis Lectures on Human Language Technologies 5 (1): 1–167.

Lopez, Jose A. 2001. "Evaluating the Predictive Accuracy of Volatility Models." Journal of Forecasting 20 (2): 87–109.

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." The Journal of Finance 66 (1): 35–65.

———. 2016. "Textual Analysis in Accounting and Finance: A Survey." Journal of Accounting Research 54 (4): 1187–1230.

Lucca, David O, and Francesco Trebbi. 2009. "Measuring Central Bank Communication: An Automated Approach with Application to Fomc Statements." National Bureau of Economic Research.

Ludvigson, Sydney C. 2004. "Consumer Confidence and Consumer Spending." Journal of Economic Perspectives 18 (2): 29–50.

Ludvigson, Sydney C, and Charles Steindel. 1998. How Important Is the Stock Market Effect on Consumption? Vol. 9821. Federal Reserve Bank of New York New York.

Manela, Asaf, and Alan Moreira. 2017. "News Implied Volatility and Disaster Concerns." Journal of Financial Economics 123 (1): 137–62.

Mankiw, N Gregory, and Ricardo Reis. 2002. "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." The Quarterly Journal of Economics 117 (4): 1295–1328.

Meinshausen, Nicolai. 2006. "Quantile Regression Forests." Journal of Machine Learning Research 7 (Jun): 983–99.

Mergner, Sascha, and Jan Bulla. 2008. "Time-Varying Beta Risk of Pan-European Industry Portfolios: A Comparison of Alternative Modeling Techniques." The European Journal of Finance 14 (8): 771–802. `http://ideas.repec.org/a/taf/eurjfi/v14y2008i8p771-802.html`.

Mishkin, Frederic S. 2004. "Can Central Bank Transparency Go Too Far?" National Bureau of Economic Research.

Mohammad, Saif M., and Peter D. Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon" 29 (3): 436–65.

Montero, Pablo, and JosÃľ A Vilar. 2014. "TSclust: An R Package for Time Series Clustering." Journal of Statistical Software 62 (1): 1–43.

Nadeau, Richard, Richard G Niemi, David P Fan, and Timothy Amato. 1999. "Elite Economic Forecasts, Economic News, Mass Economic Judgments, and Presidential Approval." The Journal of Politics 61 (1): 109–35.

Ndou, Eliphas, Nombulelo Gumata, and Mthuli Ncube. 2017. "Business Confidence Shocks and the Relevance of Exchange Rate Volatility and Economic Policy Uncertainty Channels." In Global Economic Uncertainties and Exchange Rate Shocks, 383–99. Springer.

Nielsen, Finn ÃĚrup. 2011. "Afinn." Richard Petersens Plads, Building 321.

Odendaal, Nicolaas Johannes, Monique Reid, and Johann F Kirsten. 2018. "Media Based Sentiment Indices as an Alternative Measure of Consumer Confidence."

Ooms, Jeroen. 2017. Pdftools: Text Extraction, Rendering and Converting of Pdf Documents. `https://CRAN.R-project.org/package=pdftools`.

Peng, Yangtuo, and Hui Jiang. 2015. "Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks." arXiv Preprint arXiv:1506.07220.

Phillips, Peter CB, and Pierre Perron. 1988. "Testing for a Unit Root in Time Series Regression." Biometrika 75 (2): 335–46.

Pigou, Arthur Cecil. 1927. "Industrial Fluctuations."

Price, S McKay, James S Doran, David R Peterson, and Barbara A Bliss. 2012. "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone." Journal of Banking & Finance 36 (4): 992–1011.

Prollochs, Nicolas, Stefan Feuerriegel, and Dirk Neumann. 2015. "Generating Domain-Specific Dictionaries Using Bayesian Learning." In ECIS.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. `http://www.R-project.org/`.

Reid, Monique, and Stan Du Plessis. 2011. "Talking to the Inattentive Public: How the Media Translates the Reserve Bank's Communications." Department of Economics and the Bureau for Economic Research at the University of Stellenbosch.

Romer, Paul. 2016. "The Trouble with Macroeconomics." The American Economist 20: 1–20.

Said, Said E, and David A Dickey. 1984. "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order." Biometrika 71 (3): 599–607.

Santero, Teresa, and Niels Westerlund. 1996. "Confidence Indicators and Their Relationship to Changes in Economic Activity."

Shapiro, Adam Hale, Moritz Sudhof, and Daniel Wilson. 2017. "Measuring News Sentiment." In. Federal Reserve Bank of San Francisco.

———. 2018. "Measuring News Sentiment." In. Federal Reserve Bank of San Francisco.

Shapiro, Harold T. 1972. "The Index of Consumer Sentiment and Economic Forecasting: A Reappraisal." Human Behavior in Economic Affairs, 373–96.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." JOSS 1 (3). `https://doi.org/10.21105/joss.00037`.

Sims, Christopher A. 2003. "Implications of Rational Inattention." Journal of Monetary Economics 50 (3): 665–90.

Sokal, Robert R, and F James Rohlf. 1962. "The Comparison of Dendrograms by Objective Methods." Taxon 11 (2): 33–40.

Soroka, Stuart N, Dominik A Stecula, and Christopher Wlezien. 2015. "It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion." American Journal of Political Science 59 (2): 457–74.

Souleles, Nicholas S. 2004. "Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys." Journal of Money, Credit, and Banking 36 (1): 39–72.

Stephens-Davidowitz, Seth. 2014. "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data." Journal of Public Economics 118: 26–40.

Stock, James H., and Mark W. Watson. 2007. Introduction to Econometrics Boston: Pearson Addison Wesley.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." BMC Bioinformatics 8 (1): 25.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. "Extracting Semantic Orientations of Words Using Spin Model." In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 133–40. Association for Computational Linguistics.

Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." The Journal of Finance 62 (3): 1139–68.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." The Journal of Finance 63 (3): 1437–67.

Theil, Christoph Kilian, Sanja Stajner, and Heiner Stuckenschmidt. 2018. "Word Embeddings-Based Uncertainty Detection in Financial Disclosures." In Proceedings of the First Workshop on Economics and Natural Language Processing, 32–37.

Toda, Hiro Y, and Taku Yamamoto. 1995. "Statistical Inference in Vector Autoregressions with Possibly Integrated Processes." Journal of Econometrics 66 (1-2): 225–50.

Trapletti, Adrian, and Kurt Hornik. 2019. Tseries: Time Series Analysis and Computational Finance. `https://CRAN.R-project.org/package=tseries`.

Tyralis, Hristos, and Georgia Papacharalampous. 2017. "Variable Selection in Time Series Forecasting Using Random Forests." Algorithms 10 (4): 114.

Van den Brakel, Jan, Emily SÃűhler, Piet Daas, and Bart Buelens. 2017. "Social Media as a Data Source for Official Statistics; the Dutch Consumer Confidence Index." Survey Methodology 43 (2).

Venter, JC. 2019. "The Sarb's Composite Business Cycle Indicators." In Business Cycles in Brics, 425–46. Springer.

Wallis, Kenneth F. 1986. "Forecasting with an Econometric Model: The 'Ragged Edge'problem." Journal of Forecasting 5 (1): 1–13.

Ward Jr, Joe H. 1963. "Hierarchical Grouping to Optimize an Objective Function." Journal of the American Statistical Association 58 (301): 236–44.

Weigend, Andreas S, and Shanming Shi. 2000. "Predicting Daily Probability Distributions of S&P500 Returns." Journal of Forecasting 19 (4): 375–92.

West, Kenneth D. 1996. "Asymptotic Inference About Predictive Ability." Econometrica: Journal of the Econometric Society, 1067–84.

Wickham, Hadley. 2017. Tidyverse: Easily Install and Load the 'Tidyverse'. `https://CRAN.R-project.org/package=tidyverse`.

Wilcox, James A. 2007. "Forecasting Components of Consumption with Components of Consumer Sentiment." Business Economics 42 (4): 22–32.

Woodford, Michael. 2005. "Central Bank Communication and Policy Effectiveness." National Bureau of Economic Research.

Wright, Marvin N, and Andreas Ziegler. 2015. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." arXiv Preprint arXiv:1508.04409.

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." Political Communication 29 (2): 205–31.

Zellner, Arnold. 1971. "An Introduction to Bayesian Inference in Econometrics."