Media based sentiment indices as an alternative measure of consumer confidence

HANJO.ODEN@GMAIL.COM ODENDAAL, NICOLAAS JOHANNES
MREID@SUN.AC.ZA REID, MONIQUE

Stellenbosch Economic Working Papers: WP17/2018

www.ekon.sun.ac.za/wpapers/2018/wp172018

September 2018

KEYWORDS: Big Data, Sentiment Analysis, Consumer Confidence
JEL: B41, C52, C55, C83

Bureau of Economic Research (BER)
www.ber.ac.za

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

Media based sentiment indices as an alternative measure of consumer confidence

Nicolaas Johannes Odendaal[a], Monique Reid[b]

[a]*Stellenbosch University, South Africa; Bureau of Economic Research, South Africa*

[b]*Stellenbosch University, South Africa;*

Abstract

The world is currently generating data at an uprecedented rate. Embracing the data revolution, case studies on the construction of alternative consumer confidence indices using large text datasets have started to make its way into the academic literature. These 'sentiment indices' are constructed using text-based analysis. A subfield within computational linguistics. In this paper we consider the feasibility of constructing online sentiment indices using large amounts of media data as an alternative for the conventional survey method in South Africa. A clustering framework is adopted to provide an indication of feasible cadidate sentiment indices that best reflect the traditional survey based confidence consumer index conducted by the BER. The results indicate that the best candidate indices are linked to a single data source with a focus on using specialised financial dictionaries. Finally, composite indices for consumer confidence is constructed using Principle Component Analysis. The resulting indices' high correlation with the traditional consumer confidence index provide motivation for using media data sources to track consumer confidence within an emerging market such as South Africa using sentiment based techniques

Keywords: Big Data, Sentiment Analysis, Consumer Confidence, Emerging Markets

JEL classification B41, C52, C55, C83

1. Introduction

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 from physical sampling or surveying. The private sector has researched 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. As a consequence of these developments, economists are starting to embrace the data revolution by rethinking how this kind of data could be employed within standard 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

*Corresponding author: Nicolaas Johannes Odendaal
Email addresses: hanjo.oden@gmail.com (Nicolaas Johannes Odendaal), mreid@sun.ac.za (Monique Reid)
**Acknowledgement of funding*: This research was supported by the South African Bureau of Economic Research (BER). We thank our colleagues from BER who provided insight and expertise that greatly assisted the research.

collection on a large scale was used within economics to complement more traditional methods is the Billion Price Project of Cavallo (2013). Under the BBP, prices from hundreds of online stores, spanning over fifty countries, are collected daily. In this way, it offers an alternative to traditional price indices that has the advantages of being a real time measure that is also available at a far higher frequency.

Macroeconomic data is usually released with a substantial delay, complicating the task of forecasting and making economic decisions within both policy institutions and the private sector. One of the primary tasks of market and policy orientated institutions, as well as economists, are creating nowcasting estimates. Many of these institutions dedicate substantial effort to nowcasting - the econometric techniques to predict the current (now) state of the economy as well as the that of the recent past. Accurately predicting the current economic condition or key policy indicators of the economy, such as GDP growth, is hampered by revisions, asynchronous explanatory variables and ragged-edge issues. The ragged edge problem relates to the asynchronous manner in which economic statistics becomes available within a given period, e.g., month or quarter.

To address some of these issues, nowcasting techniques have focused on integrating high frequency information predictors that aren't subject to subsequent data revisions. Some have focused on econometric methods that can model mixed frequency data and combine model forecasts (Thorsrud and others (2016)). Nowcasting models generally consists of both 'hard' and 'soft' information (Shapiro, Sudhof, and Wilson (2017)). Hard predictors are objective, quantifiable variables, which include economic values of production, employment and demand, while soft information captures subjective variables typically collected through responses from surveys. The main objective of the consumer confidence survey is to capture populace attitude towards the current and future economic climate. Despite the extensive and voluminous literature on nowcasting, central banks still have a hard time capturing the underlying economic state when it changes rapidly. Alessi et al. (2014) illustrates the extent of this problem by examining evidence from the Great Recession - a time when fast and accurate forecasting performance was considered an invaluable commodity for economists.

This paper investigates online news media, which is publically available source that possibly contains information about the state of the economy and consumer's attitudes towards it. We aim to contribute to the literature in a number of ways. Firstly, the paper 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[1]. The survey-based business and consumer confidence indices developed by the Bureau of Economic Research of South Africa (BER) are the most widely quoted in South Africa. Their indices are based on the famous confidence index of the University of Michigan (UM). Research has confirmed that the confidence index developed by 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

[1]For countries with a diverse socio-economic landscape, it could be argumented that the opinion in the news is not a proxy attitude for the population as a whole

(Souleles (2004), Bram and Ludvigson (1997), Curtin (2007)).

A second contribution of the paper is to suggest a framework by which researchers and practitioners can develop a monthly index that could act as an alternative to the traditional survey based consumer sentiment index[2] using multiple online media channels and dictionaries. These indices can be used as complementary or alternative indices within national statistics as an indicator for consumer sentiment. By incorporating daily news and editorial content, the index aims to capture market information from a plethora of different channels. These channels form the opinion of not only professional, but a shared view of economic agents (as represented by the authors of the media articles). This unstructured information-set offers higher dimensionality and higher frequency than survey based methods. The data and approach allows for the analysis of economic fluctuations through a bottom up modeling approach. A time series clustering technique is be 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 CCI. The aim is to identify the data sources and dictionaries which best mimics the traditional survey based index.

The rest of this paper is structured as follows. An overview of the current literature on consumer confidence and how to measure it is presented in section 2. This is followed, section 3, with an explanation of how computational linguistics, and the "Bag-of-Words" approach, can be used to extract sentiment from a piece of unstructured text. Section 4 gives an extensive overview of the data. In section 5, we present the methodology adopted to construct the indices from the and a description of how smoothing techniques are employed to deal 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 a final media based sentiment index (MSI) that is constructed from a principle 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.

2. 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

[2]Sentiment analysis forms part of a larger field called computational linguistic and although consumer confidence and sentiment can be thought of as interchangeable concepts, in this paper sentiment refers to the constructed sentiment score or polarity as derived by computational analysis

> 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 forth 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) explores the causes of the 1990-1991 recession, by trying to isolate the causal sentiment variable[3]. Angeletos and La'O (2013) formalise 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 equilibrium. The authors relax the assumption that economic agents should have expectations that match the actual state of the economy and in doing so, illustrates 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 is derived by Benhabib, Wang, and Wen (2015). They employ 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 (Benhabib, Wang, and Wen (2015)).

The second channel through which it is proposed that consumer confidence affectes the business cycle, is an information contagion channel. 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. 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 affect 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) find 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 the main argument in times of economic expansion and "animal spirits" in times of contraction.

For both the 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.

[3]Causal alluding to an economic variable which @blanchard1993consumption states to be a proximate cause or decrease in relation to its normal determinants

2.1. Survey method for measuring consumer confidence

The concept of consumer confidence originated in the mid 1940s with George Katona at the University of Michigan. The traditional approach to measuring consumer confidence is to construct a sentiment 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 as to have a quantitative manner 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[4].

In South Africa, a consumer confidence survey is conducted on a quarterly basis by the Bureau of Economic Research of South Africa (BER)[5]. 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. In terms of the majority of the population group, white and black, the 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, 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[6].

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 paper only focusses 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

[4]This entails subtracting the percentage negative responses from the favourable answers. The index is based to have an index value of 100 in December 1964
[5]The BER is the only institution in South Africa that conducts consumer confidence surveys on a regular basis
[6]The validation check is done either in person or telephonically

conducting qualitative assessment of the business environment using the German-based Ifo method back in March 1954[7]. 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.

3. Sentiment through textual analysis

Computational linguistics is best known for its sentiment and topic analysis tool-set[8]. 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; and arousal describes and measures the intensity of the emotion.

Recent case studies on the construction of consumer confidence indices using online media data have been appearing in 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 (S. C. Ludvigson (2004)). However the indices suffer from the disadvantage of not knowing the exact sampling population. They also face a substantial opportunity cost in constructing a fully automated electronic pipeline to analyse all of the text information.

In addition, there are choices that need to be made when constructing the media-based sentiment indices which was 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. 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

[7] 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

[8] To read up on how topic analysis is being applied within the field of economics and finance see @hansen2016shocking, @larsen2015value or @hansen2014transparency

used Harvard Psychosoiological Dictionary (Harvard IV), is due to negative connotations of words like tax, capital, expenditure, risk etc which are neutral within a financial context. The dictionaries being used are:

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

Table 3.1 illustrates the 'Bag-of-words' concept by showcasing associated word and its related emotion from 5 different dictionaries which will be used in the analysis section of this paper.

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

Table 3.1: *Example of associated words and emotions from five different dictionaries*

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 to text analysis employs machine learning algorithms within what is called 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 consists 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. Liu (n.d.) examines in detail the effect of subjectivity in the estimation of sentiment within a piece of text. 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 the sentiment scores could be different across different models.

One of the first papers to investigate the feasibility of a constructing a social media sentiment index, using the "Bag-of-words" method, was P. J. 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 concluded that a strong association does exist between consumer confidence (measured through surveys) and the online sentiment (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, is driven by the same underlying phenomenon.

Brakel et al. (2017) further investigates concept of using social media to derive sentiment. They analyse a Dutch SMI by using a multivariate structural time series approach to estimate whether the inclusion of social media in the production of Dutch administrative statistics improve their accuracy. The paper also investigates 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 is 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 modeling procedure as seen in Harvey and Chung (2000), Brakel et al. (2017)'s results indicate that the Dutch CCI and SMI do indeed follow a similar evolutionary process, albeit that what the underlying data generation different.

Fraiberger (2016) turn to news information to infer economic sentiment. Fraiberger (2016) uses 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 a 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 over time. Dictionary methods can also be used to capture uncertainty and is not just used within the context of sentiment analysis. Baker, Bloom, and Davis (2016) develops 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 paper, 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, while already NLP scored sentiment data was acquired from Meltwater[9]

4. Data

In order to construct an economic news sentiment index, we collected data from three different sources: Meltwater, Sabinet and News24. These sources represent a large readership throughout South Africa with the aim of capturing a general sentiment of the diverse readership found in emerging markets such as South Africa. As a reference to what is generally considered to be the golden standard for confidence in South Africa, we obtain data from the BER that contains the current consumer confidence (CCI) which is constructed through surveys. These surveys will be used to compare our constructed news sentiment index to. The analysis restricts itself to the period February 2009 to October 2017 as all data-sets have information for this time frame.

The data provided by Meltwater is 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 bolean search criteria was then exported to a CSV file at an aggregated level. 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-dived into 3 categories:

- Consumer confidence
- Business confidence
- Job-market

The keywords used in searching the editorial are aimed at extracting as much information as possible on the current economic environment[10].

[9] a company specialising in NLP technology and online sentiment analysis. Insight into who Meltater is and what they do can be found here: https://www.meltwater.com/

[10] The full Boolean query used to extract each of the three data samples is available on request from the authors

Figure 4.1 shows the different sources and the number of articles identified through the boolean searches. In total, 207 314 articles were assigned a sentiment score. Meltwater's data-set is unique in this analysis, as the sentiment scores for the articles were already assigned without any intervention from the authors.

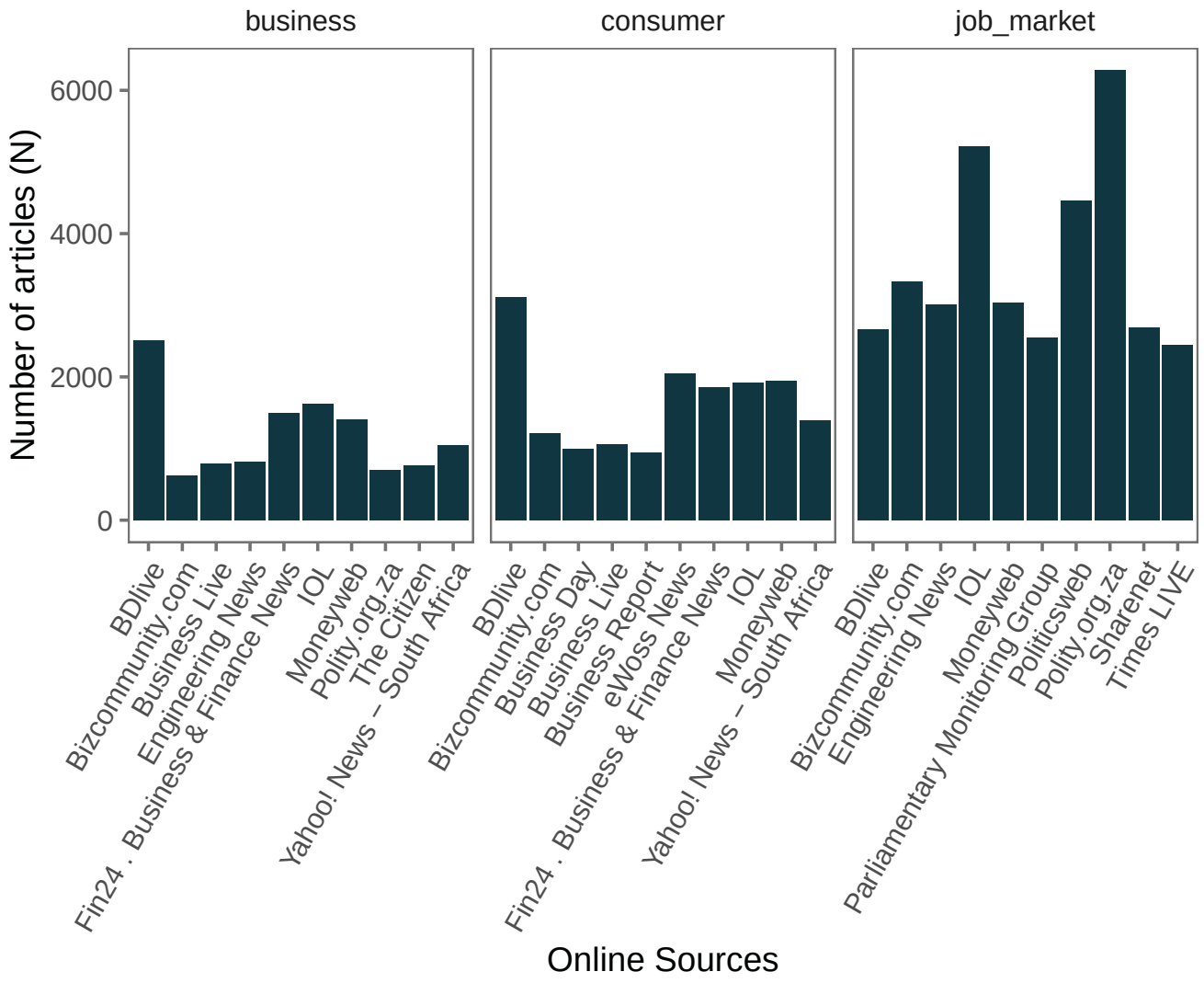

Figure 4.1: *Sources of articles as identified thorugh boolean searches*

Our second source for news articles came from News24. The data received was in raw text format. This allowed for a much deeper analysis of the text. The data-set 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 the English articles received as 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 further, 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 data-set of 271 000 articles, of which approximately 70 000 articles were labelled "financial". Other labels consisted out of "archived", "local" or "South African". Figure 4.2 gives an indication of the number of articles being published online over the period of investigation.

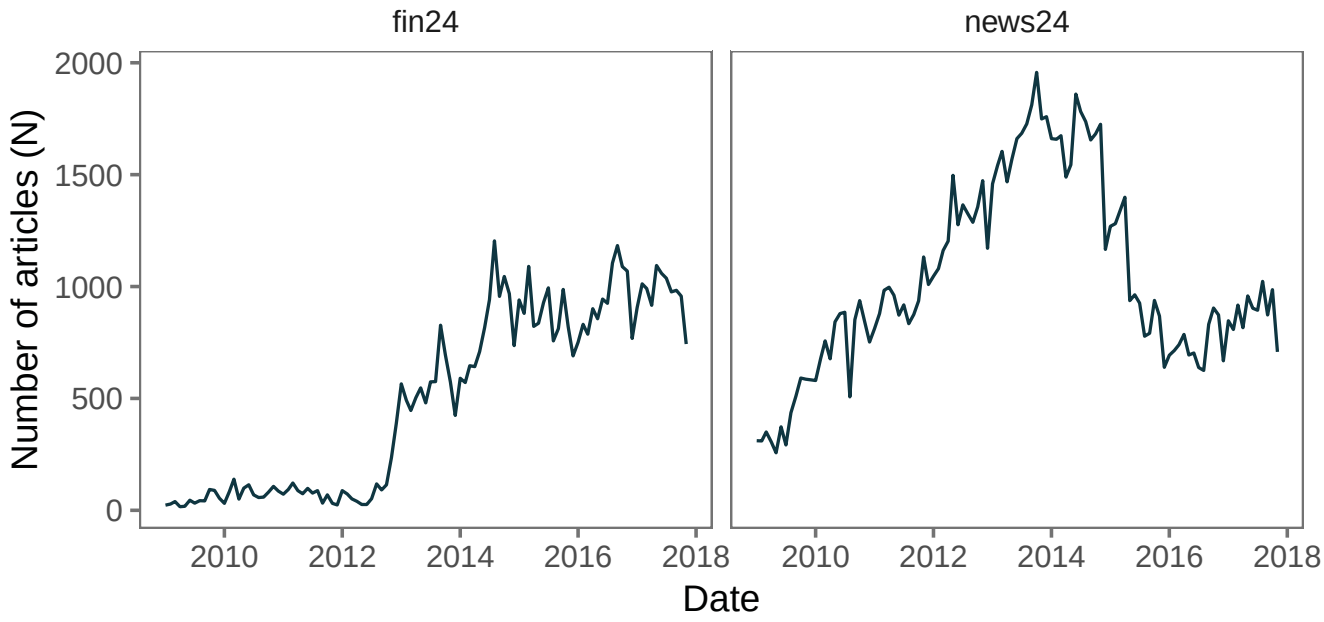

Figure 4.2: *Number of articles from January 2009 to October 2017*

It is clear that for articles with a "financial" tag, the bulk of the corpus exists after 2012. This is mainly due to News24 moving towards online platforms and promoting the availability of the published materials online counterparts for a low subscription fee. A majority of the articles were classified under a tag called "archive". Given that not much is known about these articles, it was decided that these should fall under the general "News24" banner, rather than try and extract the financial news. The final count of news articles for analysis was 163 526.

The final source of text data was Sabinet. The dataset consisted out of 262 300 PDF files from three different news sources:

- Business day (160 000)
- Financial Mail (34 000)
- Sowetan (68 300)

These newspapers represent a large readership throughout South Africa. The aim of using these specific publishers is to capture general sentiment through diverse readership. The Sowetan is by far the biggest newspaper of out of the three, having an exclusive readership of around 500 000. The Business Day can be considered the most read newspaper aimed at financial matters and has a readership of approximately 80 000. The smallest of the three, the Financial Mail, is a very concentrated newspaper with a key focus in the financial, investment and political space. The Financial Mail's readership hovers around 40 000. All of these newspaper 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 then 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 were 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 discarded[11]. Another problem with this data source, was missing observations for the period 2015-01-01 to 2015-05-01. In order to accommodate the construction of the indices, the data for the missing period was substituted with data from the "News24" data-set.

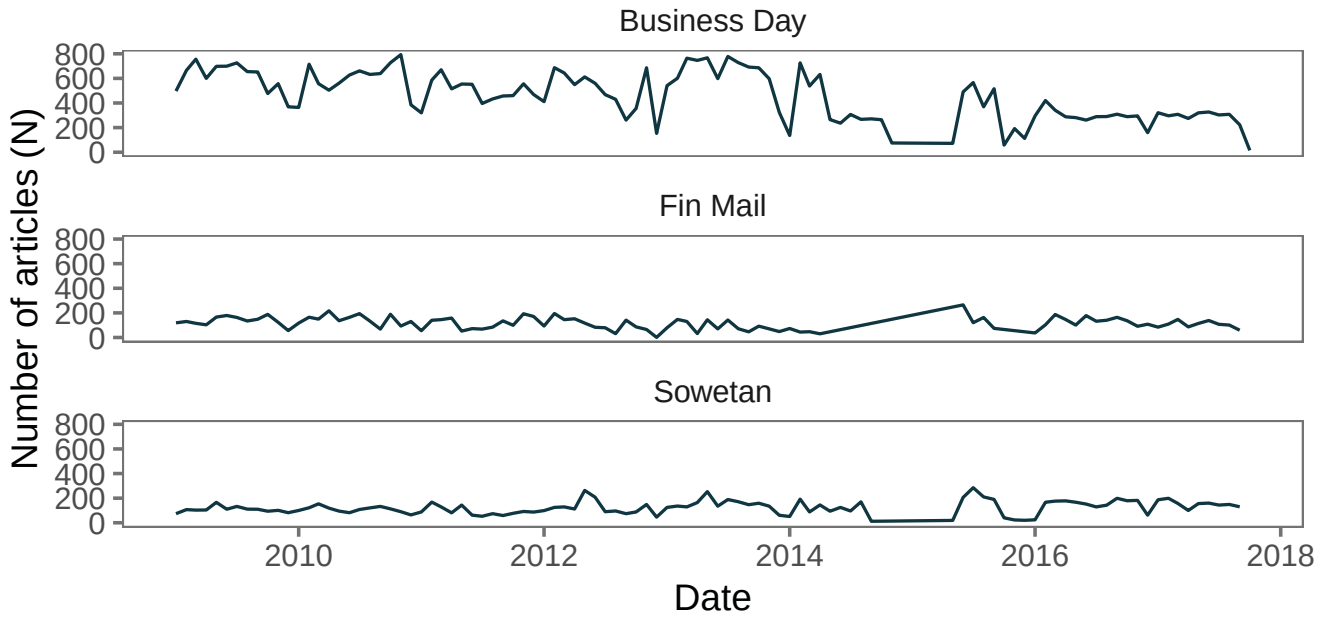

Figure 4.3: *Number of articles from January 2009 to October 2017 as provided by Sabinet*

The missing articles can be seen in the figure 4.3 as straight lines between the points for the period beginning 2015.

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

5. Methodology

5.1. Creating a sentiment score

The total articles for a given data-set N_i^a where $i \in \{Meltwater_x, Sabinet_x, News_x\}$ and x represents the samples within each data-set, while the total period over which the analysis will be conducted is represented by $T^d = \{2009\text{-}02\text{-}01{:}2017\text{-}09\text{-}31\}$. As is standard with any analysis conducted on a corpus of this size, a data preparation step is introduced before analysis can begin.

The Meltwater data-set already contains an associated sentiment score per article and thus does 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) and the `tidytext` library by Julia Silge[12].

The first step of cleaning raw text data is to remove all stop words. Stop words are commonly used words such as *the*, *and*, *a* and so forth, that do not contribute anything towards our understanding of the content of the underlying text. The lexicon we employ 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 paper, we do not stem words, as the dictionaries that we will be using do not require this. The last bit of cleaning involves removing all editorial pieces with less than 150 words as not to bias the dictionary method with low word count articles.

To construct a sentiment score, we identify the positive and negative words in each article $N_i^a \quad \forall \quad T$ using an external word lists (dictionary), and doing a simple word count that consists of the positive plus negative words. We then normalise the count so that it reflects the relative proportion of positive and negative words within an article:

$$Pos_{i,t,n^a} = \frac{PositiveWords}{PositiveWords + NegativeWords} \quad Neg_{t,n^a} = \frac{NegativeWords}{PositiveWords + NegativeWords} \tag{5.1}$$

where $i \in \{Sabinet_x, News_x\}$ [13]. 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:

[12] Other packages we use as part of the data cleaning forms part of what is known as the tidyverse (@wickham2017)
[13] Meltwater already providing a sentiment per article

$$S_{i,t,n^a} = Pos_{i,t,n^a} - Neg_{t,n^a} \tag{5.2}$$

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 dataset now recoreds the Data Provider, ID, Date, sentiment score and polarity of each of news report. The index is constructed as the net balance of positive and negative articles within a month. This is the same method the University of Michigan uses to construct their well known consumer confidence index.

We create five different indices for each of our raw text data-sets using each dictionary. Once the sub-sample indices are created we also construct source level indices as an arithmetic average from the sub samples. Figure 5.1 provides a visual representation of all the data-sets used in the construction in the indices.

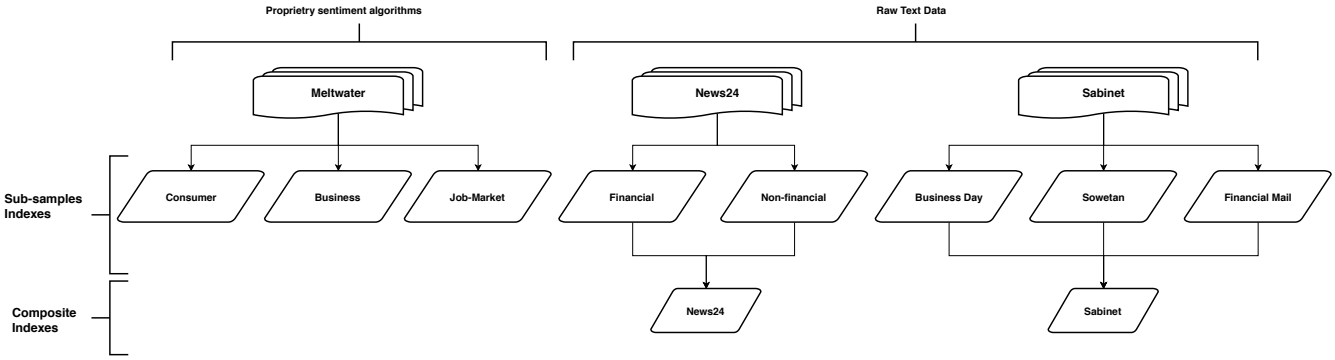

Figure 5.1: *Full diagram of data-sets used in analysis*

5.2. Smoothing the sentiment index

Due to the volatile nature of the monthly indices, all series are smoothed using a Gaussian local level model[14]. 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 $t = 1, \ldots, T$:

$$\xi_{t+1} = T_t \xi_t + c_t + R_t \eta_t \quad \eta \sim N(0, Q_t) \tag{5.3}$$

$$y_t = Z_t \xi_t + d_t + \epsilon_t \quad \epsilon_t \sim N(0, H_t) \tag{5.4}$$

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

Equations (5.3) and (5.4) 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 modeled. 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 paper, the latter drift vectors, $\mathbf{c}_t$ and $\mathbf{d}_t$, will be set to zero, to only capture the mean component.

The positive definite covariance matrices, $\mathbf{Q}_t$ and $\mathbf{H}_t$, are in fact the serially uncorrelated, normally distributed $r \times 1$ and $N \times 1$ vector error terms, η_t and ϵ_t, with mean zero:

$$E(\eta_t \eta_\tau') = \begin{cases} \mathbf{Q}_t & \text{for } t = \tau, \\ \mathbf{0} & otherwise \end{cases} \tag{5.5}$$

$$E(\epsilon_t \epsilon_\tau') = \begin{cases} \mathbf{H}_t & \text{for } t = \tau, \\ \mathbf{0} & otherwise \end{cases} \tag{5.6}$$

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

$$E(\eta_\tau \epsilon_t') = 0 \quad \text{for all } \tau, t, \ldots, T \tag{5.7}$$

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 $\mathbf{P}_1$:

$$\xi_1 \sim N(a_1, P_1) \tag{5.8}$$

The local level model is analogous to a 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[15] will be used throughout. We believe a deeper discussion and exploration into the

[15]This implies $a_1 = 0$ and $P_1 \to \infty$, as proposed in @durbinkoopman.

wide ranging choice of priors is not warranted given the scope of this paper and recommend 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 characterized by its mean, $\mathbf{a}_{t+1}$, and its covariance, $\mathbf{P}_{t+1}$:

$$\mathbf{a}_{t+1} = E(\xi_{t+1}|\mathbf{Y}_t) \tag{5.9}$$

$$\mathbf{P}_{t+1} = Var(\xi_{t+1}|\mathbf{Y}_t) \tag{5.10}$$

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 $\mathbf{P}_t$[16] 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:

$$\hat{\xi}_t = E(\xi_t|y) \tag{5.11}$$

$$V_t = Var(\xi_t|y) \tag{5.12}$$

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.

We can see how the different data providers and dictionaries make a big difference when comparing the constructed

[16]For a full mathematical derivation of this, see @mergner.

indices to the benchmark BER CCI in figure 5.2 and 5.3. The figure for the Meltwater data is seperated from the raw text sources due to the nature of construction.

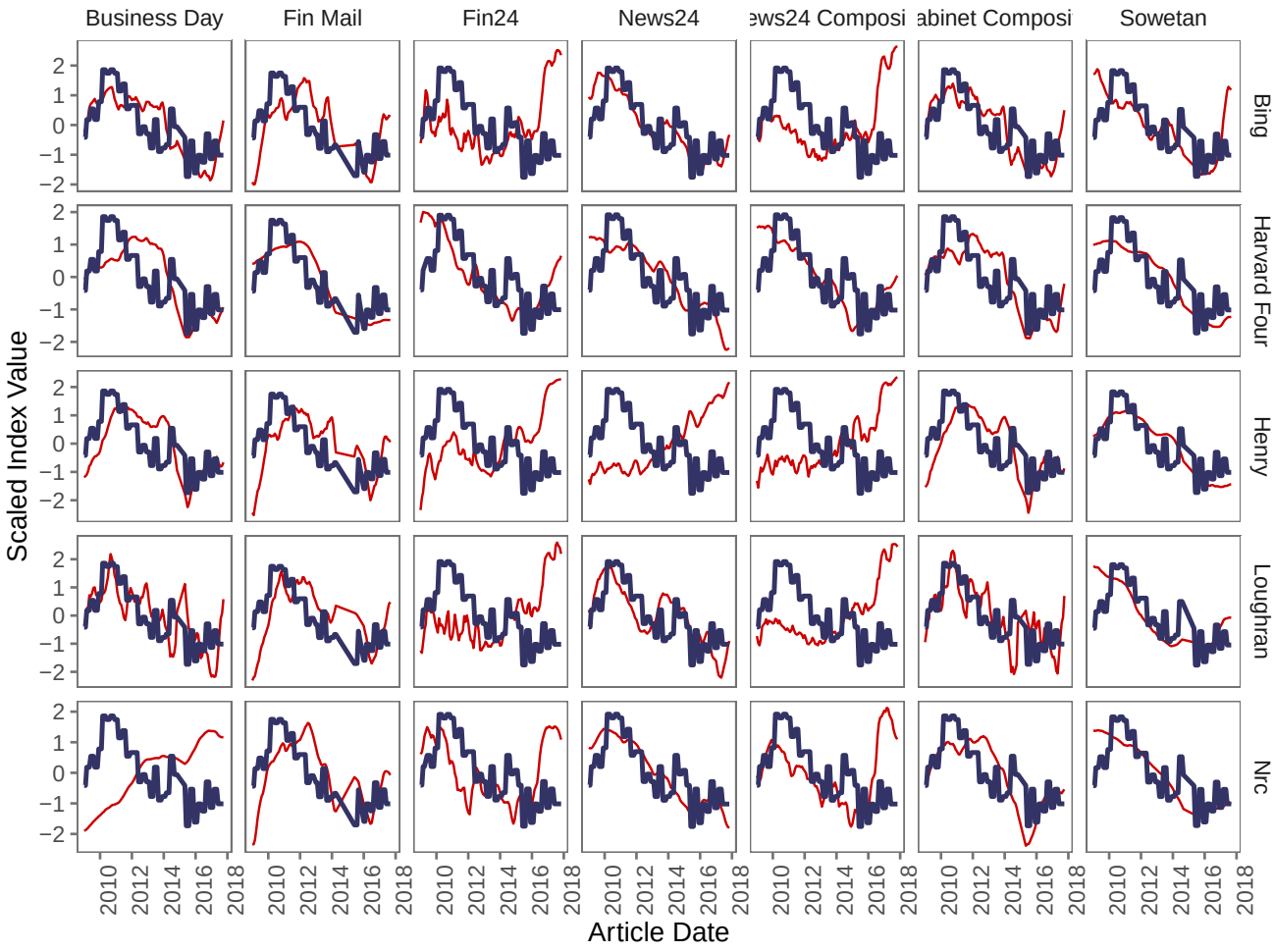

Figure 5.2: *All indices constructed from boolean search and different dictionaries overlayed onto the CCI of the BER*

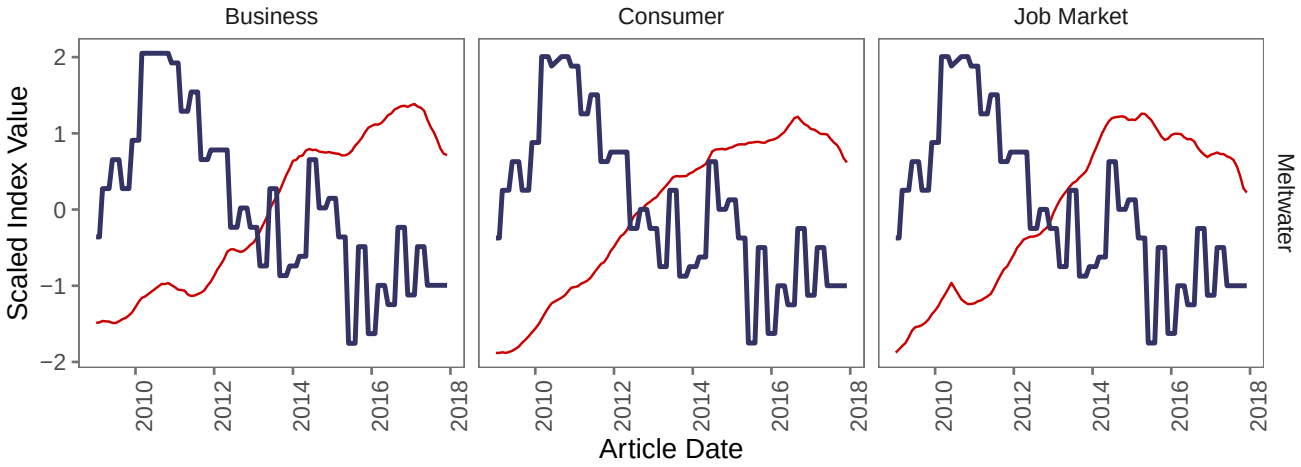

Figure 5.3: *Meltwater indices constructed from boolean search overlayed onto the CCI of the BER*

5.3. Quantitative analysis

Once all the indices have been constructed we use 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 analysis section of the paper aims to identify the best confidence candidates that mimic's the BER's indicator[17] from the numerous indices created in the previous section. To accomplish this, we turn to the field of time series clustering methods. The aim is to sub-divide the large sample set, through clustering, into smaller homogeneous buckets to identify which of the indices created resemble the lowest dissimilarity with the BER's consumer confidence index.

Time series clustering is an active research area with applications of the techniques being seen in literature encompassing a wide range of fields. Although the technique is gaining traction within other fields, the technique is still under utilised within the field of economics. We employ this technique in order to help us identify time series that behaves in the same way. The hypothesis is that series that move similar to the CCI would cluster in a homogenous group. The steps in applying the analysis is two-fold. The first step is to identify an appropriate dissimilarity matrix between the indices created, the CCI. Secondly we use 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, Vilar, and others (2014)). Each of these methods have their own strength and weakness and as such, a "best" dissimilarity measure will differ with

[17]The CCI is considered to be current benchmark indices in South Africa to measure consumer confidence

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 around comparing geometric profiles of a series, while structure-based dissimilarity constructs aim to compare underlying data generating processes (or structures).

For our study its important that the direction of change be the same and because of this, we decide 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 dont need to be the same. This paper also settles on this dissimilarity measure, as distance measures such as pearson does 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. The problem the analysis faces when choosing an appropriate distance metric is the time domain of the series we have to cluster. 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[18].

We start off defining r as a proximity measure and M representing all possible sequence of m pairs where we preserve the ordering of the observations by imposing *monotonicity* while also avoiding futile looping:

$$r = ((X_{a_1}, X_{b_1}), \ldots, (X_{a_m}, X_{b_m})) \tag{5.13}$$

where $ai, bj \in \{1, \ldots, T\}$ such that $a_1 = b_1 = 1, a_m = b_m = T$, with the distance between the coupled observations (X_{ai}, X_{bi}) is minimised.

$$d_{DTW}(\mathbf{X_T}, \mathbf{Y_T}) = \min_{r \in M} \left(\max_{i=1,\ldots,m} |X_{a_i} - Y_{b_i}| \right) \tag{5.14}$$

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 example of how dynamic time warping creates a mapping between time series can be seen in figure 5.4 (Giorgino and others (2009)):

[18] For an extensive list on more complex dissimilarity measures, see @montero2014tsclust

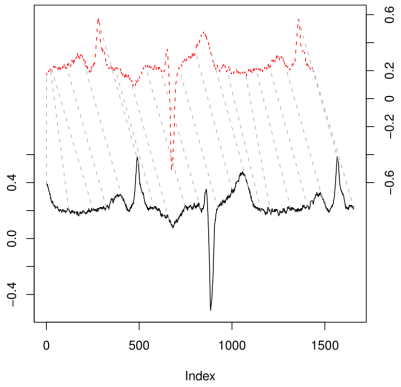

Figure 5.4: *Illustration of how Dynamic Time Warping is a mapping function between two points from* (X_{ai}, X_{bi})

Table 5.1 shows the result of the top 10 closest constructed online sentiment indices as per the dynamic time warping metric:

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

Table 5.1: *Distance between CCI and the constructed media sentiment indices per DTW*

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 produces time series which best reflects 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 know 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

then repeated until a final cluster is formed in a tree-like fashion. Both paradigms of hierarchical clustering posses what is known as a monotinicity property. This suggests that the 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. This paper will use Ward Jr (1963)'s method which calculates a merging cost when forming clusters A and B. Let $A = \{a_1, \ldots, a_{n_a}\}$ and $B = \{b_1, \ldots, b_{n_b}\}$ consist out of observations in $\mathbb{R}^d$. Define the between-within, or e-distance $e(A, B)$, between A and B as:

$$e(A,B) = \frac{n_A n_B}{n_A + n_B}\left(\frac{2}{n_A 1 n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(a_i, b_j)\right. \tag{5.15}$$

$$\left. - \frac{1}{n_A^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_A} d(a_i, a_j) - \frac{1}{n_B^2}\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} d(b_i, b_j)\right). \tag{5.16}$$

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 points in the cluster is minimised.

5.4. Clustering results

Figure 5.5 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, by Hubert and Levin (1976). This resulted in the CCI, Business Day (Loughran) and Sabinet Composite (Loughran) ending up in a single cluster.

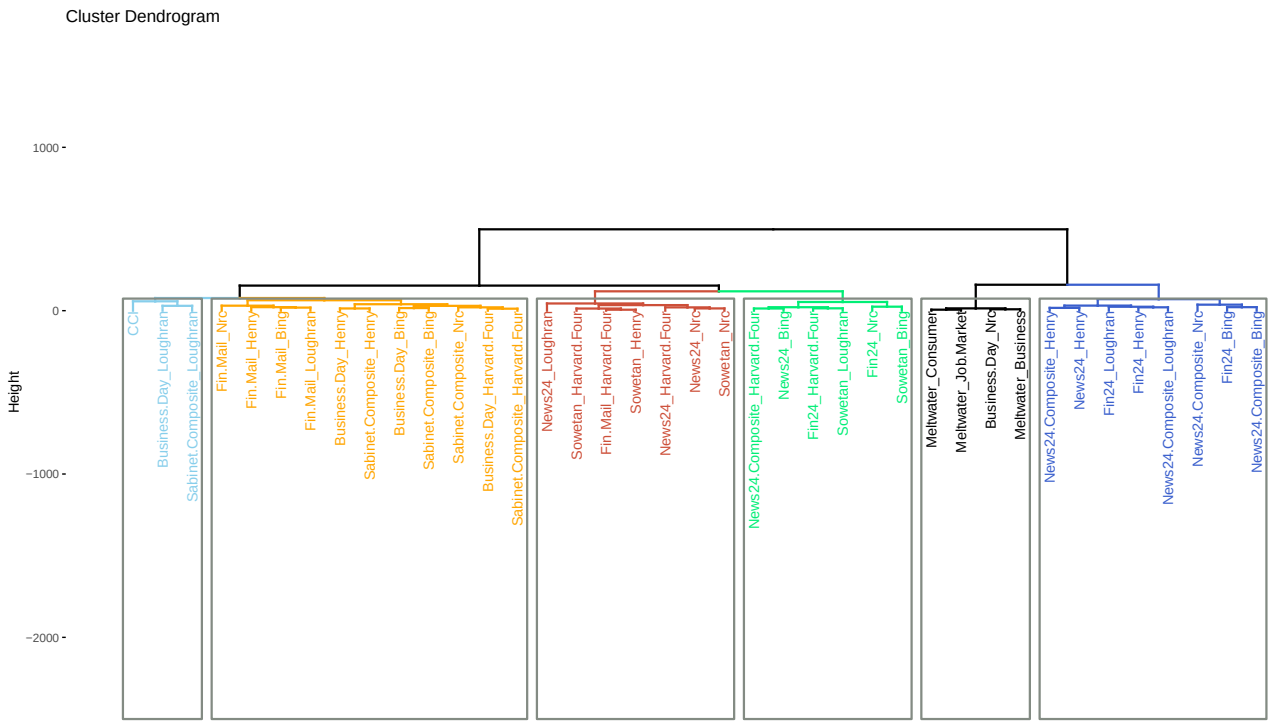

Figure 5.5: *Visualization of the hclust results*

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 and financially orientated lexicons deliver the best results[19]. Focusing on the cluster that contains the CCI, three data sources provide similar indices: Sabinet, Business Day and Fin Mail. 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 with this specific editorial focus that is most likely the reason this group of Fin Mail is observed in the CCI cluster. All of these constructed indices reiterate the important role the financial dictionaries are playing in constructing a series that assimilates 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 results in quite a smooth series, while Sabinet - of which Business Day is a subset - exhibits much larger changes between points in the series.

[19] 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

6. Discussion

The results from the clustering analysis reveal that there are clusters of indices that mimic the current survey method CCI, even though the constructed online sentiment indices are all monthly and the survey methods are quarterly. This is an interesting finding on its own, as the news indices have no specific sampling that would allow it to capture the diverse population groups within the country. Despite this disadvantage that the online media has in this regard, the clustering approach does well in capturing indices that resemble the CCI. These results lead to the construction of three seperate media sentiment indices (MSI) for evaluation. The results of the clustering algorithm and the raw DTW measure is used as input to construct two composite indices, while the series most similar to the CCI, based on DTW, is also considered.

To construct the composite media based consumer sentiment indices (MSI), we use principle component analysis. Principle component analysis can be seen as using a set 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 6.1 shows the result of the constructed indices, from the PCA, for the two MSIs as well as the single index MSI. The figure gives an indication of how the different constructions of a media based index compare to the BER's CCI.

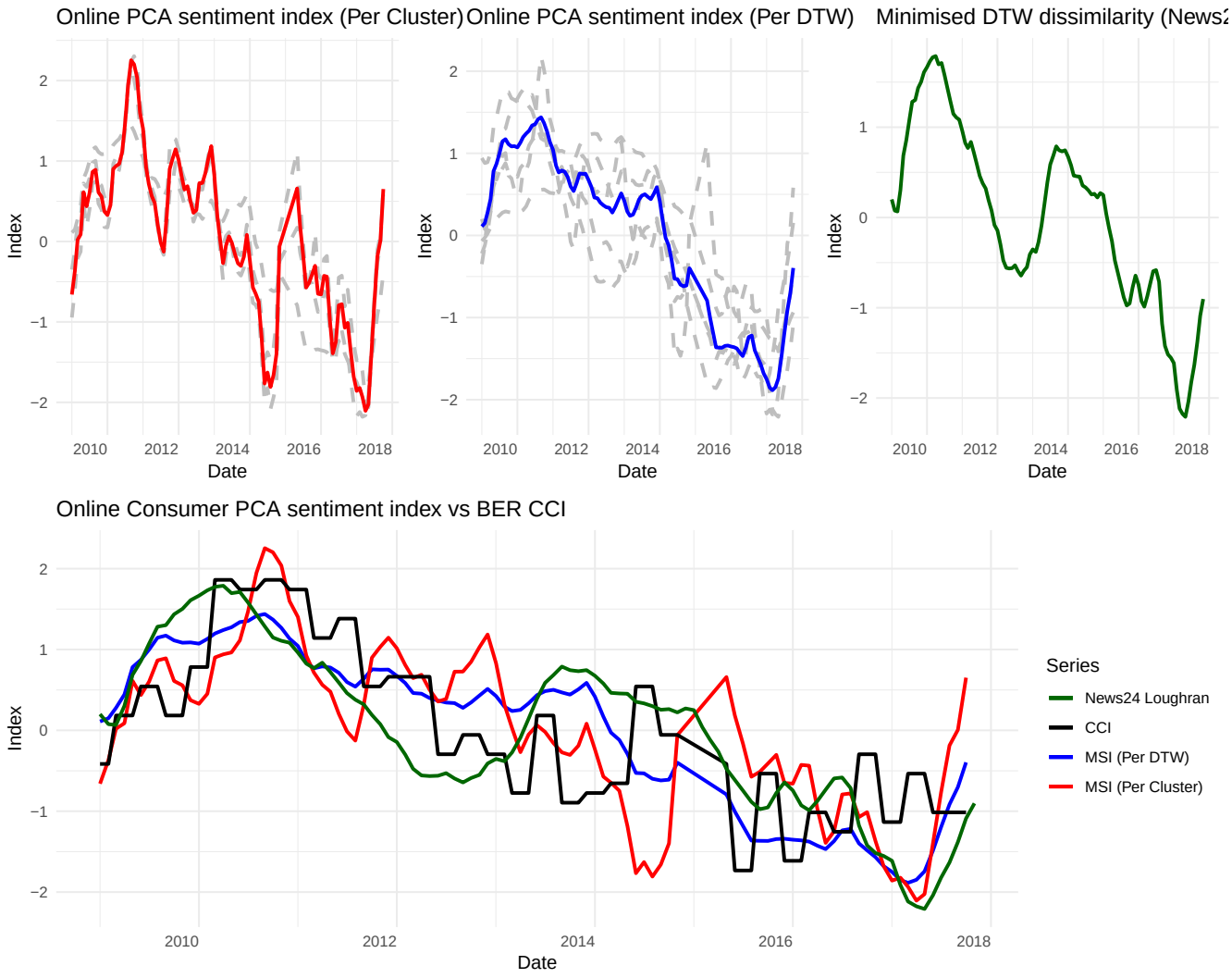

Figure 6.1: *(top left) The results of the PCA on series found in the CCI cluster. Red line represents the scaled first dimension of the constructed online sentiment indicex.(top middle) The results of the PCA on based on the dissimilarity measure. The blue line represents index. (top right) The single index that minimized the distance between the CCI and itself as per DTW (Bottom) The results of the 3 media sentiment indices overlayed on top of the BER CCI (black).*

Table 6.1 presents the PCA results of the composite online consumer sentiment indices. The results from the PCA show that the first component captures the majority of the variation present in collection of time series, with a cumulative percentage of variance equal to 96.29% and 78.40%. The construction of the indices are also not skewed to a single series, with each series' contribution to the first vector being almost.

	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 used to construct 2 different MSI indices based on the clustering results as well as just using the top 5 indices most similar to CCI (as per DTW)*

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 2 constructed MSI indices*

Table 6.1: *Composite consumer index PCA results*

To help answer the question of whether the indices can be considered alternatives for the traditional CCI, the indices need to either have a high contemporaneous correlation or lead the index. In order to construct the correlation between the monthly and quarterly series we convert the monthly indices to a quarterly one. This is done as correlation is a function of variance and covariance between two series and keeping the CCI as a step function will bias the correlation result. Table 6.2 indicates the correlation of the different MSI measures with the CCI as well as the lagging component when we lag the MSIs from t to $t-4$.

Index	t	t-1	t-2	t-3	t-4
MSI (Per Cluster)	0.58	0.56	0.66	0.60	0.53
MSI (Per DTW)	0.76	0.75	0.79	0.75	0.68
News24 Loughran	0.72	0.75	0.77	0.77	0.67

Table 6.2: *Correlation between CCI and the constructed media sentiment (and their lags from t to t-4). Correlation measure was calcualted on quarterly data.*

The highest correlation in the table is obtained when we lag the MSI (per DTW) with $t-2$. This would translate

to the MSI series leading the CCI with 2 quarters. The MSI (per DTW) also has the highest mean correlation with CCI over the different lags. The second highest correlation was observed by the raw News24 Loughran index. This index had a mean above 0.70 across the timeframes. The MSI (per Cluster) also sees an increase in the correlation with the CCI in the $t-2$ and $t-3$ periods, but not as strong as the other two series under investigation.

All the indices show promise as alternative measures of sentiment in the consumer market. However the cluster method does seem to provide an index that does not fit the CCI as well as the DTW measures. To confirm the validity of these series as economic indicators, further research needs to be conducted on whether the series is able to predict future consumer and business activity. Another topic of interest would be to see if the series provides informational content above and beyond the current measures of confidence. Refinement in the construction of the indices also needs to be investigated, as this paper aims to only suggest a generalised framework for the investigation of high dimensional data to track economic sentiment. Especially the work on statistically deciding on how many series should be included in the construction of an MSI series using raw DTW values.

7. 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 as well as provide an insight into likely current economic conditions. This paper's aim was to investigate the feasibility of constructing online sentiment indices using various different corpora. We suggest the use of a clustering framework to select the most appropriate data sources and lexicon dictionaries to apply within the "bag-of-word" approach. As a main goal, the paper aimed to investigate whether an online based sentiment index resembles a survey based approached. If a mildly representative index is constructed, it has the advantage over the currently implemented BER confidence indices, which is only released quarterly and needs to have field surveys 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 was 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 and large amounts of online editorial data that does resemble the BER's current confidence index. The different MSIs shows a high level of correlation with the BER's CCI, ranging between $\rho = 0.58$ and $\rho = 0.79$ for different series and timeframes. These results give way to future research in the field of constructing confidence indices using alternative data sources. The findings of this paper motives for the further investigation of the use of large text data to construct indices that

could be alternatives for the traditional survey based methods.

References

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. Elsevier: 86–97.

Alessi, Lucia, Eric Ghysels, Luca Onorante, Richard Peach, and Simon Potter. 2014. "Central Bank Macroeconomic Forecasting During the Global Financial Crisis: The European Central Bank and Federal Reserve Bank of New York Experiences." *Journal of Business & Economic Statistics* 32 (4). Taylor & Francis: 483–500.

Angeletos, George-Marios, and Jennifer La'O. 2013. "Sentiments." *Econometrica* 81 (2). Wiley Online Library: 739–79.

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

Barsky, Robert B, and Eric R Sims. 2011. "News Shocks and Business Cycles." *Journal of Monetary Economics* 58 (3). Elsevier: 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. 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). Wiley Online Library: 549–85.

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.

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

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

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

Sentiment Index Horse Race."

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

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

Curtin, Richard. 2007. "Consumer Sentiment Surveys: Worldwide Review and Assessment." *OECD Journal. Journal of Business Cycle Measurement and Analysis* 2007 (1). Organisation for Economic Cooperation; Development (OECD): 7.

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

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

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

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

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). Wiley Online Library: 303–9.

Henry, Elaine. 2008. "Are Investors Influenced by How Earnings Press Releases Are Written?" *The Journal of Business Communication (1973)* 45 (4). SAGE Publications Sage CA: Los Angeles, CA: 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). American Psychological Association: 1072.

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).

MIT Press: 209–23.

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.

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. n.d. "Sentiment Analysis and Subjectivity." 627–66.

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). Wiley Online Library: 35–65.

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

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

Montero, Pablo, José A Vilar, and others. 2014. "TSclust: An R Package for Time Series Clustering." *Journal of Statistical Software* 62 (1). Foundation for Open Access Statistics: 1–43.

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

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

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

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). The Ohio State University Press: 39–72.

Thorsrud, Leif Anders, and others. 2016. "Nowcasting Using News Topics. Big Data Versus Big Bank."

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

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts."

Political Communication 29 (2). Taylor & Francis: 205–31.