

Media based sentiment indices as an alternative measure of consumer confidence

Evidence for South Africa

Nicolaas Johannes (Hanjo) Odendaal

Department of Economics/ BER

<https://www.daeconomist.com>



What is the news really trying to say?

More information is always better than less. When people know the reason things are happening, even if it's bad news, they can adjust their expectations and react accordingly.

Research question

The research attempts to gauge the feasibility of constructing online sentiment indices using large amounts of text data as an alternative for the conventional survey method

- Can news wholly act as an indicator on consumer confidence
 - Analysis of economic fluctuations through a bottom up modeling approach
 - We construct a monthly series to try and explore this hypothesis
- We illustrate and suggest a quantitative framework which researchers and practitioners could employ to help build their own indices
 - This is especially important when the data is of a very high-dimensional nature ¹

[1] As is the case with text documents

Understanding consumer confidence

Animal spirits

The 'animal spirits' hypothesis was a thought experiment first put forth by Keynes 1937

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)

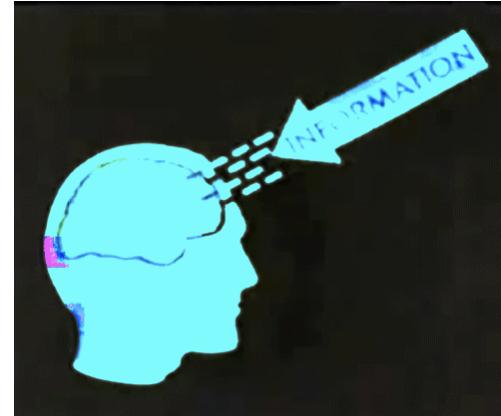
- An unexpected change in the business cycle could occur due to the 'gut', or sentiment outlook, of economic agents reacting out of subjective foresight, rather than *quantitative* evidence
- This in turn changes economic activity through a consumer sentiment shock

There is also a *contradicting* philosophy where informational news is the main driving component

Informational contagion

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 amount of unexpected business cycle fluctuations can be attributed towards 'animal spirits'; stating that uncaptured fundamental news is the primary channel by which the relationship of sentiment and subsequent economic activity exists



From the two above explored philosophies, our research sides with information contagion and hypothesize that that informational news about the economy could be exploited. This is especially exciting if you can achieve this a high frequency.

Current literature on the subject

Where there is smoke there is fire...

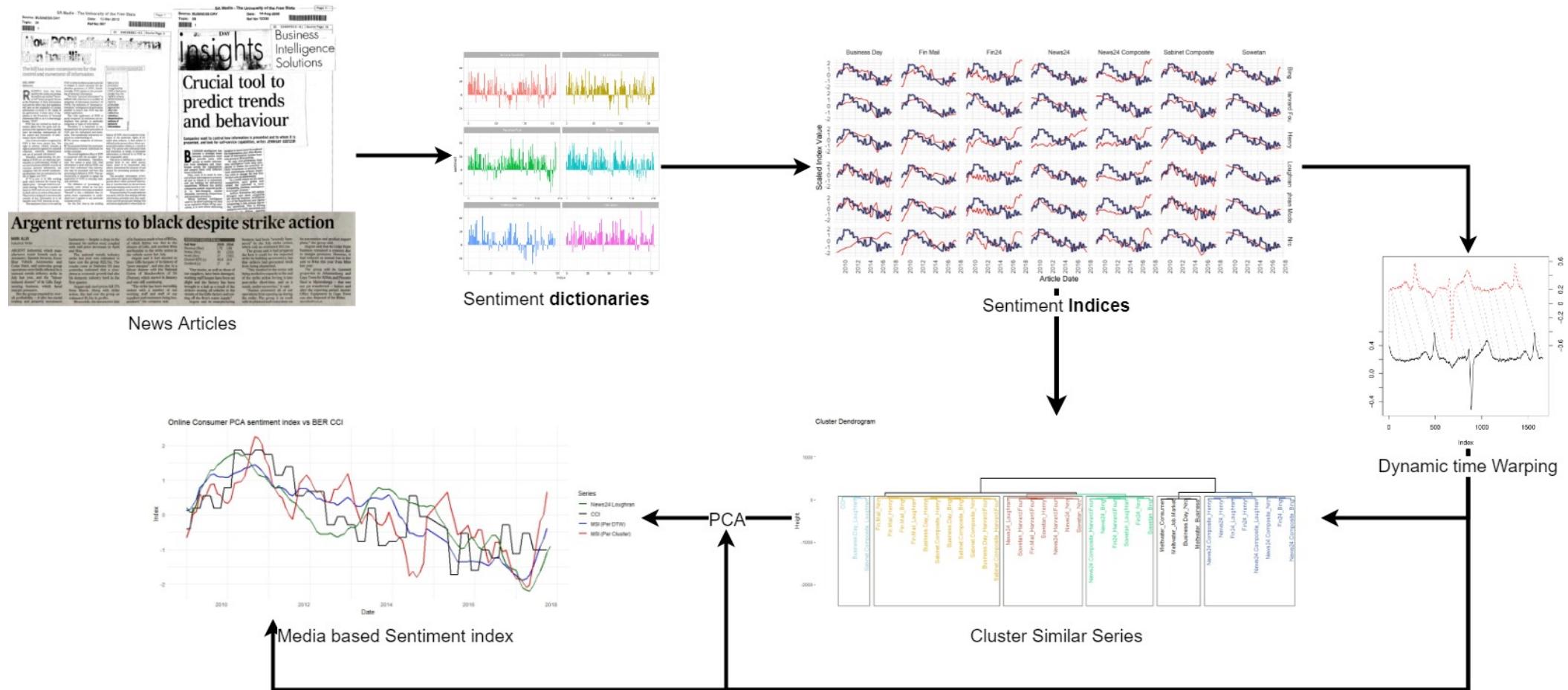
- P. J. Daas and M. Puts (2014)
 - Considers the construction of a Dutch *social media* sentiment index (SMI) derived from Facebook, Twitter and various other online data sources. The finding of the paper concluded that a strong association exists between consumer confidence and the sentiment displayed by public Facebook messages.
- Brakel et al. (2017)
 - Investigates the 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
 - Their findings motivates that the SMI can be seen as a substitution for the more traditional survey approach

The Dutch National Statistics office was one of the first to start exploring the idea ¹, but the practical implications when using social media data is disadvantageous.



[1] I have to recognise Piet Daas and Marco Puts work - they are the reason why I am doing my PhD ;-)

Proposed framework to build sentiment indices



Traditional survey approaches towards measuring consumer confidence

The consumer confidence questions consists out of the following:

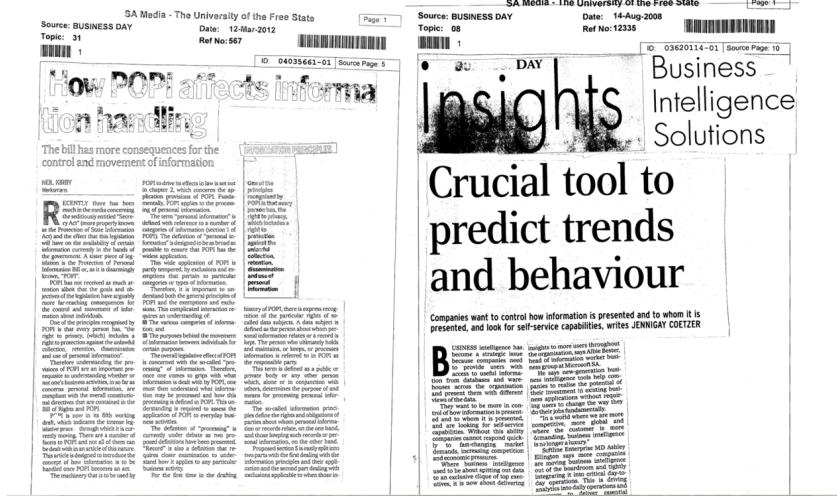
1. *How do you expect the general economic position in South Africa to develop during the next 12 months? Will it improve considerably, improve slightly, deteriorate slightly, deteriorate considerably or don't know?*
2. *How do you expect the financial position in your household to develop in the next 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?*

Index constructed as a normalised sum of relative scores.

Percentage of respondents expecting an improvement / good time less the percentage expecting a deterioration / bad time

Text-based construction of confidence indicators

A dictionary "bag-of-words" approach to sentiment mining is widely used.



Argent returns to black despite strike action

Text-based construction of confidence indicators (Cont.)

Doing a simple word count that consists out of the positive and negative words, we normalise the count so that they reflect the relative fraction of positive and negative words within an article:

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

The polarity of the article is derived from the score.

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

If the score of the article is greater than zero, then the overall sentiment for the article is deemed to be positive, the same can be said for a negative sentiment score.

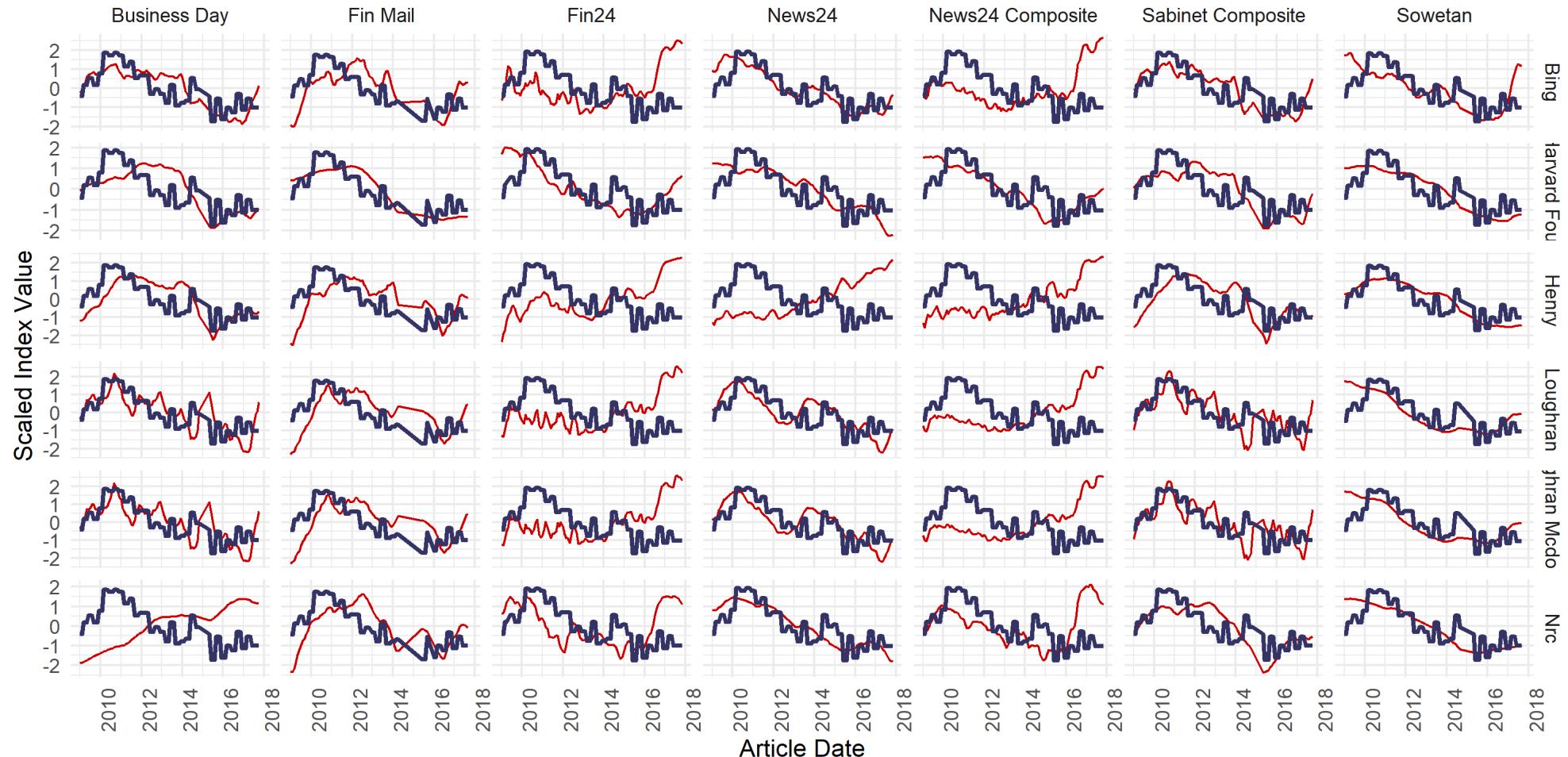
Text-based construction of confidence indicators (Cont.)

The index is constructed as the *net balance of positive and negative articles* within a month. This method of constructing an index should be familiar, as it is the same method the University of Michigan uses to construct their well known consumer confidence index

$$I_t = \frac{\text{PositiveArticles} - \text{NegativeArticles}}{\text{PositiveArticles} + \text{NegativeArticles}}$$

- All indexes are finally smoothed using a Kalman Filter

How do the indices compare against CCI?





Isolating noise from signal

The aim is to identify the best confidence candidates that mimic's the BER's CCI from the plethora of indexes created.

Isolating noise from signal

The aim is to identify the best confidence candidates that mimic's the BER's CCI from the plethora of indexes created.

- 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 indexes created resemble the lowest dissimilarity with the BER's consumer confidence index

Isolating noise from signal

The aim is to identify the best confidence candidates that mimic's the BER's CCI from the plethora of indexes created.

- 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 indexes created resemble the lowest dissimilarity with the BER's consumer confidence index

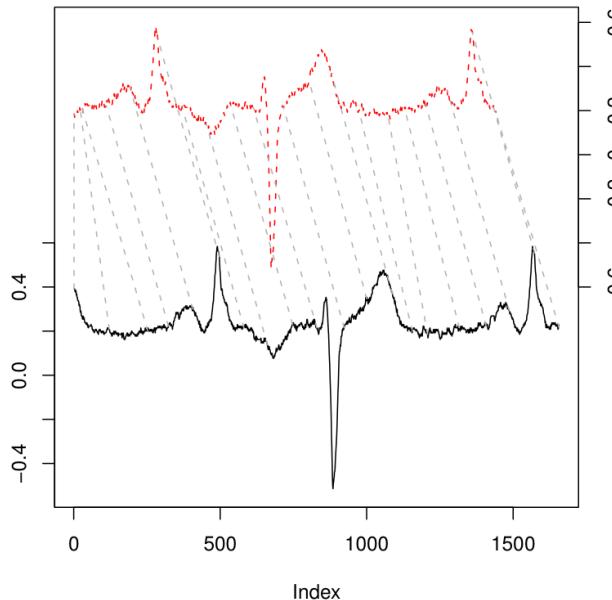
The steps in applying the analysis is two-fold

- Step 1
 - Identify an appropriate dissimilarity matrix between the indices created and the CCI
 - Use hierarchical clustering to construct clusters which consists out of all the indices
- Step 2
 - Use both the distance measure and the formed clusters to build composite indices using PCAs

From this we hope to identify a cluster which has indices similar to the traditional CCI

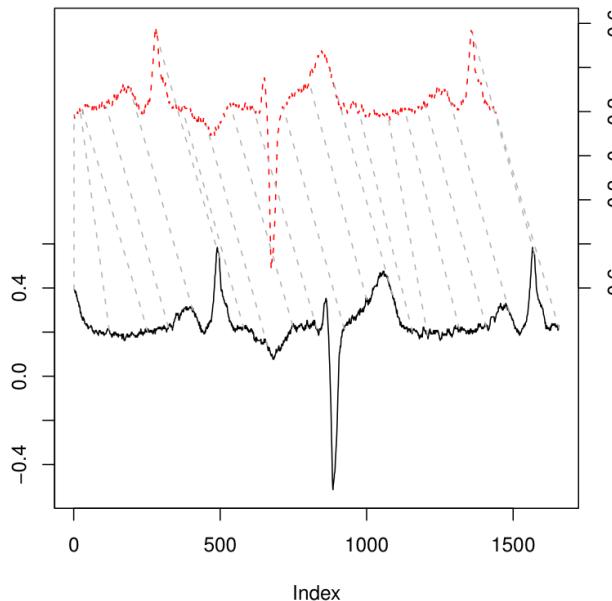
Choice of dissimilarity measure

Decide to use a non-model based dissimilarity measure, called **Dynamic Time Warping**. 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 here:



Choice of dissimilarity measure

Decide to use a non-model based dissimilarity measure, called **Dynamic Time Warping**. 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 here:



- Advantage is that a pearson correlation based measure would discount any of the constructed series which might lead the CCI!

Clustering

Hierarchical clustering has two paradigms when approaching the clustering problem, agglomerative and divisive. To conduct the clustering we use a commonly known hierarchical clustering method from `hclust` which is an agglomerative algorithm.

- The algorithm starts by assigning each observation to its own cluster
- Next it computes the similarity between each of the clusters and joining those who are most similar

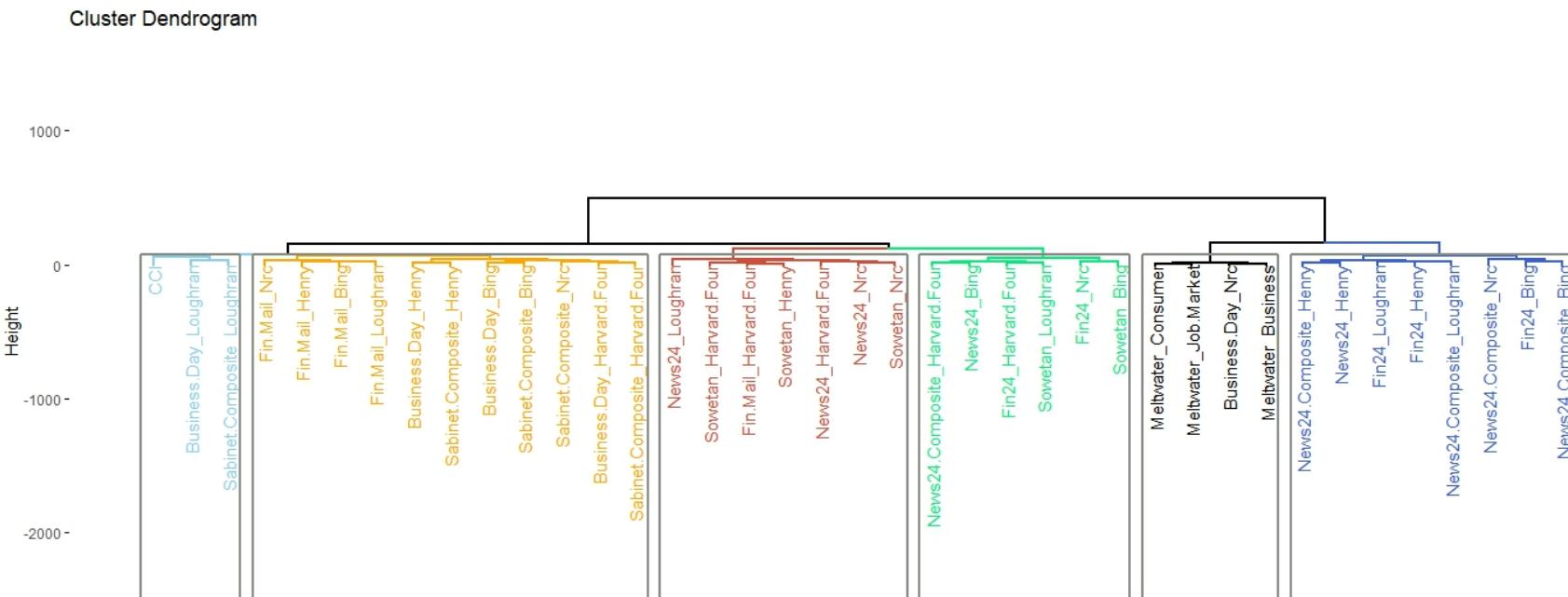
This procedure is repeated until a final cluster is formed in a tree-like fashion.

- Height of the tree at each node is proportional to the value of the inter-group dissimilarity between its daughter nodes

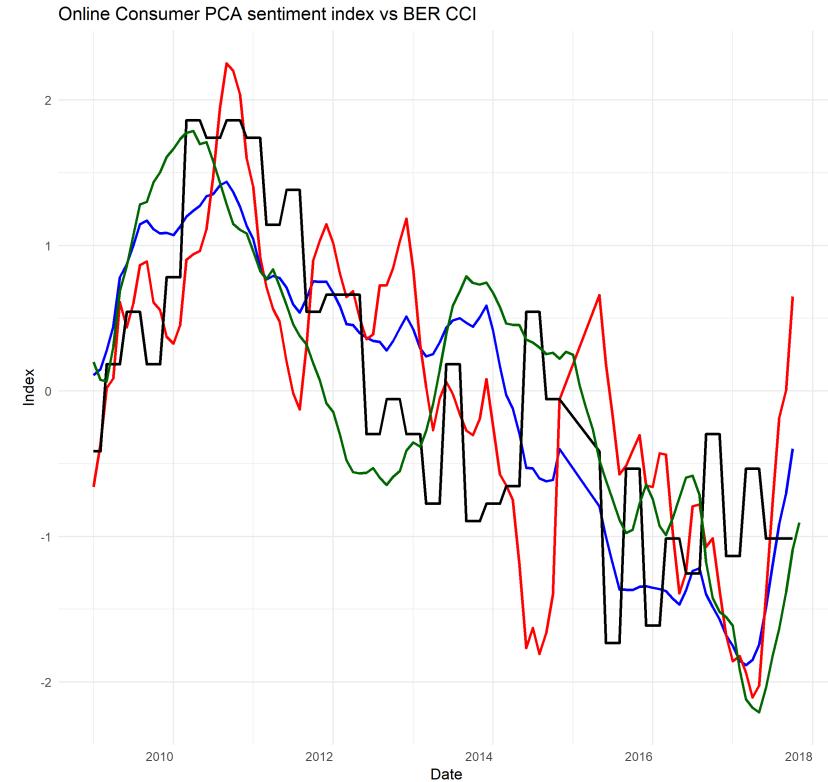
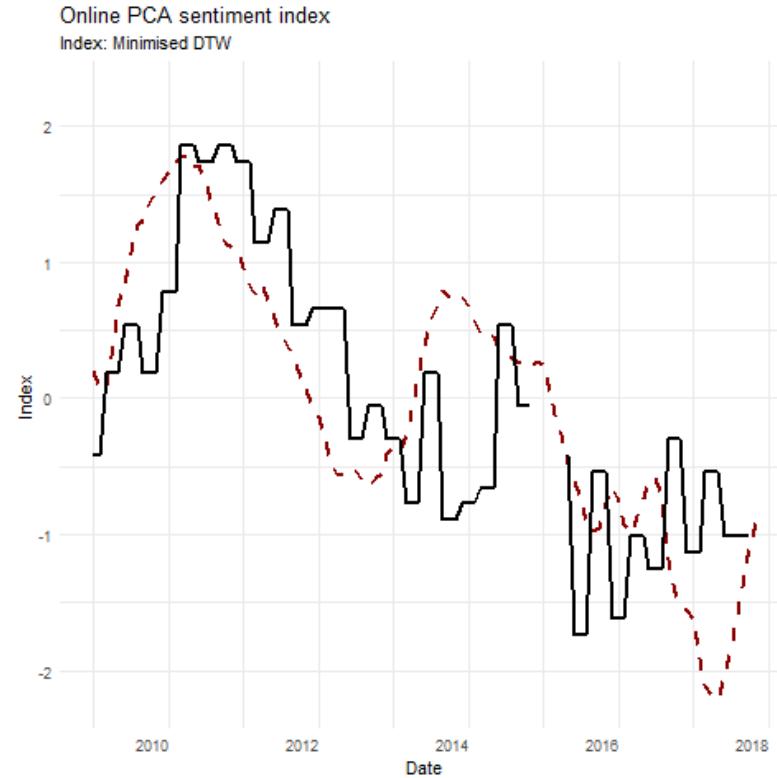
This structure is more commonly known as a *dendrogram*

Clustering: results

The graph visualises a separation between two of the clusters and the other four. The series which forms part of the most left-hand side cluster all have one thing in common, an increasing overall trend, or an increasing trend over the period of 2017:

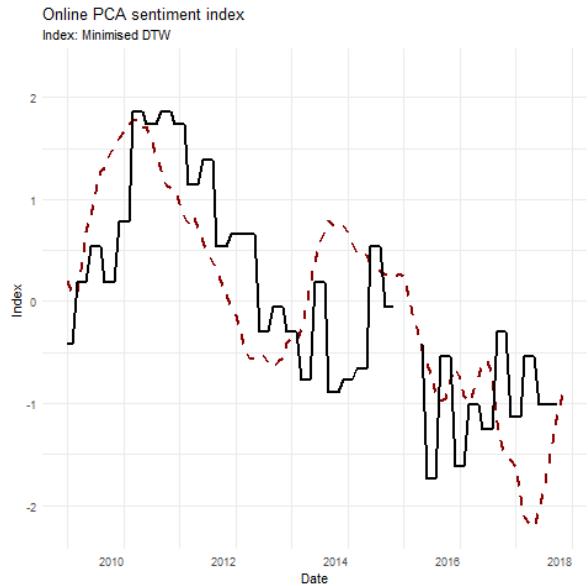


Principle component results: Consumer (Cont)



Comovement Analysis

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



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

Conclude

It is possible to construct a media based sentiment index that could potentially act as an alternative to the traditional survey based consumer confidence index.

- Developed monthly index (composite MSI) that measures economic sentiment that resembles the traditional CCI
- Motivates for the use of large text data as an alternative source to construct economic indicators

Alonside the analaysis we construct a framework that future research can build on:

- Time series clustering techniques can aid in evaluating a large number of indexes at once within high dimensional analysis

--

R Package - Analytics: tidytext, hunspell, SnowballC, pdftools, dlm, tbl2xts, TSclust, tidyverse
(especially purrr)

R Package - Reporting: Rmarkdown, xaringan, xtable, knitr, ggfortify, FactoMineR, gganimate

Future research

To confirm the validity of these series as economic indicators, further research needs to be conducted:

- Whether the series predicts future consumer and business activity
 - Does the MSI outperform the traditional survey based CCI
 - Do they provide informational content above and beyond the current measures of confidence
 - Extended timeframes
- Refinement in the construction of the composite index
 - Use of topic models to control better for information that contributes to the index
 - Could structural topic modeling help us get an idea of structural stability of the index
 - Using the DTW measure alone. is there an optimal threshold one could potentially use; i.e cumulative percentage explained we see in the PCA

Questions