Using media sentiment to forecast daily stock returns

MASTER THESIS

JESS HJULGAARD HOLSTEIN NIELSEN

SUPERVISOR: LASSE BORK

AALBORG UNIVERSITY (2020)

Abstract

This thesis seeks to examine the relation between sentiment extracted from different text sources and the daily returns of the Dow Jones Industrial Index. The focus will be on creating daily sentiment indices from freely available text sources. Since these sources often are limited and the text data scares, large amount of data will be aggregated in order to capture as much information as possible. The ability of these indices to forecast daily changes in the Dow Jones Industrial Index will be evaluated in-sample and out-of-sample. The primary models will be vector autoregression models. However, these will be extended by using principal component analysis, partial least squares and penalized regression.

Contents

Chapter 1

Introduction

Forecasting stock price movements have long been of great interest to economists and investors alike. A large amount of studies has been made, using a plethora of different methods, with varying degrees of success. It is summarized well by the following quote in Goyal and Welch (2008):

The literature is difficult to absorb. Different articles use different techniques, variables, and time periods. Results from articles that were written years ago, may change when more recent data is used. Some articles contradict the findings of others. Still, most readers are left with the impression that "prediction works"— though it is unclear exactly what works (Goyal and Welch, 2008, p.1456).

Return predictability studies have in the past primarily focused on the use of structured data, economic indicators and fundamentals. Due to the accessibility of larger amounts of available data and more advanced computers, an increased interest has been taken in the use of unstructured data. These includes data sources such as pictures, search patterns and especially text data.

The purpose of this thesis is to examine if sentiment extracted from financial text sources is able to improve forecasts of changes in daily stock prices of the Dow Jones Industrial Average[1].

Proxies for investor sentiment are constructed from different text sources, related to financial news. Financial news can be scraped daily from many online news outlets, it should be noted that this often violate certain news websites terms and conditions. Another problem occurring is that in order to gain access to older news articles, these often must be bought directly through third party companies. Even though a lot of financial news are freely available, these free sources often come with limitations. To circumvent these obstacles this project will primarily focus on freely available news, forums and social medias. The focus will be on aggregation of large amounts of sparse text data, compared to use of very specific columns as used in other papers.

Information stored in text data from news and social medias have gained increased attention in

[1] For the remainder of this thesis the daily changes in the Dow Jones Industrial Average will be referred to as *Dow*.

the last decade, not only in academia. New methods are rapidly being developed to mine the stored information in text data. The usages range from customer service and marketing all the way to asset pricing. One example of this being Thomson Reuters News Analytics (TRNA), this application computes the sentiment of all news articles published by Reuters. Additionally, they provide services such as social media tracking, allowing users to monitor the mood on social media for example Twitter.

Related literature

Using media sentiment analysis in financial forecasting is a relatively new and unexplored area. One of the first and most influential articles related to this topic is Tetlock (2007). Tetlock examines if financial news content can be used to predict daily changes in stock markets. He creates a media measure using the column *"Abreast of the Market"* in the period 1984-1999 from *The Wall Street Journal*. To generate the media index the *Harvard General Inquirer* is used, this divides words related to sentiment into 77 different categories. By using principal component analysis on these categories, a pessimism index is created. The created principal component loads primarily on four categories related to pessimism: Negative, weak, fail and fall. The relation between the daily changes in the Dow Jones Industrial Average(DJIA) and the pessimism measure is estimated using vector autoregression (VAR). Tetlock concludes that high amounts of pessimism on the preceding day forecasts a fall in market prices. However, Tetlock finds little predictive value in the positive words.

Garcia (2013) similarly examines the link between media sentiment and asset pricing. The timeframe stretches from 1905-2005 and uses specific columns from *The New York Times*[2]. The proxy for media sentiment is created by counting the negative and positive words in the columns and normalizing the indices by the total number of words. Garcia's results supports the relation Tetlock (2007) uncovers between media sentiment and changes in DJIA. Furthermore, he finds that the effect is more prominent during periods of economic recession, which is in line with the assumption of heightened sensitivity to news during periods of economic downturns. Contrary to Tetlock Garcia finds that positive words can be used to predict stock market returns. One of the key differences between Garcia (2013) and Tetlock (2007) is the used lexicons for the creation of the sentiment indices. Garcia uses the word lists proposed by Loughran and McDonald (2011) which are specifically created for text in financial context.

The articles described above focus on finding significant statistically relationships between media sentiment and stock market movements. They primarily perform in-sample regressions and focus on news extracted from traditional sources, furthermore they both consider time periods where the availability of news where more uniform and limited.

Other authors focus on using sentiment to improve forecast accuracy of daily changes in stock markets. Some of these publications include more unconventional text data compared to the likes of *The Wall Street Journal* and *The New York Times*. These also deviates from the standard VAR approach by using machine and deep learnings methods. Bollen et al. (2013) uses large amount of tweets from Twitter to capture the public mood and examine if a baseline deep-learning forecasting model, of daily changes in the DJIA, can be improved by incorporating this mood. Tweets from Twitter containing statements such as "i fell", "i am" etc. are collected in the period February 28, 2008 to December 19, 2008. Despite having few time observations,

[2]It should be noted that the examined column changes name many times throughout the time period

the amount of collected tweets is almost 10 million. By using Granger Causality, they test if these mood states cause daily changes in the DJIA, but only finds that a few of the variables does. Bollen et al. (2013) concludes based on their results that some moods such as calm and happy can improve the accuracy of stock market prediction models, while others have a negative impact on performance.

Mudinas et al. (2018) use text data from *Financial Times*, Twitter and Reddit with deep-learning methods to forecast stock market trends. Instead of examining general mood states, they download Tweets containing cashtags($) and stock tickers for companies in the DJIA. This approach enables them to examine individual companies in the DJIA as well. Similar to Bollen et al. (2013), they create several mood attitudes and once again only finds that a few Granger causes. However, the incorporation of some sentiments into the baseline model improves the forecast performance, while some have a negative impact. Mudinas et al. (2018) examine different datasets and time periods. They conclude that the interaction between media sentiment and stock price movements are dynamic and hard to identify.

Loughran and McDonald (2011) is a study not focused on return predictability like the above but on financial textual analysis. They find that as much as three-fourths of negative words in financial text are misclassified by general purpose lexicons such as The Harvard General inquirer. These negative words, classified in a sociological and psychological manner, are often not negative in financial and business context. To improve on the misclassifications Loughran and McDonald (2011) creates new lexicons, not only for negative words, but also for additional word categories such as positive words. The misclassifications are found by examining approximately 50,000 10-ks between 1994 and 2008.

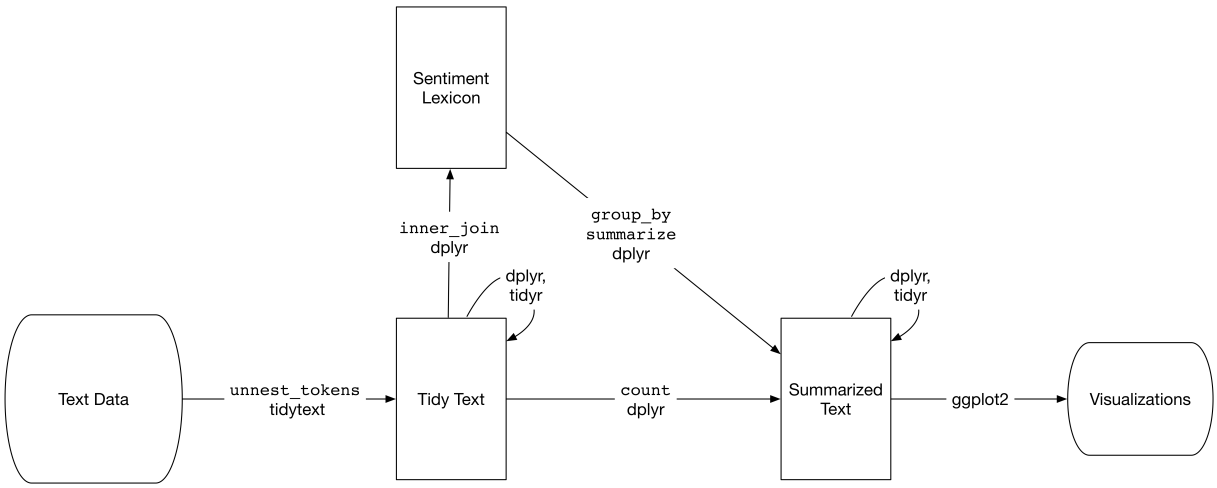
Chapter 2

Theory

2.1 Text Processing

In this project **R** has been used to perform data management, sentiment analysis, forecasting and model evaluation. Specifically, the **R** package *tidytext* has been used for the management of text data and construction of sentiment indices. The text analysis workflow is illustrated by figure 2.1.

Figure 2.1 Text analysis



Ilustration of the text analysis workflow from (Silge and Robinson, 2019, Chp.2)

Text data is downloaded and stored as character vectors in data frames. However, in order to perform an analysis on the text data, it must be tokenized. The tokenization process breaks the sentences into tokens (single words) and stores one token per row in the data frame[1]. Once this process has been completed the text data can be processed and analyzed. Firstly, stop words are removed since these are of little relevance. The most used words are counted and then grouped by day in order to construct daily sentiment indices. The text data still contains nouns and other words of little relevance, because these does not add any information about the sentiment in the

[1]This is not the entire process, punctuations are stripped and all tokens are converted to lowercase etc.

text (Silge and Robinson, 2019, chp.1). To solve this problem a lexicon approach can be used to only examine words related to the sentiment. The lexicon approach simply counts words contained within the different categories in the lexicon[2]. Since some of the used datasets in this project contains multiple text columns, such as headline and byline, these are simply combined in order to use the process described above[3].

2.1.1 Lexicons

A large number of lexicons and text analysis programs are available for sentiment analysis. One of the most well known is the Harvard General Inquirer[4], which is used in Tetlock (2007) to generate the media sentiment factor. Different versions of this lexicon exist, but the one used in Tetlock (2007) contains 77 categories. In general, the focus in financial research have been on the Harvard General Inquirer IV-4 negative and positive categories. However, since the publication Tetlock (2007) the primary focus has been on the negative category because little evidence was found of predictive power in the positive categories. Loughran and McDonald (2011) finds that negative words are often misclassified in financial context. They create a new word-dictionary containing not only negative words but also e.g. positive words. This might be the reason that Garcia (2013) finds that positive words also have a statistically influence on the *Dow*.

There exist a lot of different general-purpose lexicons. In this project the following freely available lexicons will be used:

- *bing* lexicon converts words into binary negative or positive.
- *AFINN* lexicon rates words with a score between -5 and 5, depending on how negative/positive the word is.
- *Loughran and McDonald* lexicon which was created with focus on financial text, since the general purpose lexicons often mislabels words in financial context.

The primary focus will be on the *Loughran and McDonald* lexicon due to dealing with financial data and text data related to financial news. However, the *bing* lexicon will be used to compare sample statistics[5]. The *AFINN* lexicon will be included to examine if the magnitude of the positive/negative words have any predictive information. It should be noted that the *AFINN* lexicon is a general-purpose lexicon and might not be able to capture the nuances in financial texts. Additionally these methods and lexicons does not take preceding words and the structure of the text into consideration, thus it is not able to capture negated text and irony (Silge and Robinson, 2019, chp. 2.1). An example of this in (Tetlock, 2007, p.1167) demonstrates the problem with lexicon approach. The sentences "No, the economy is not strong" and "It is not that the economy is not strong" have opposite meanings. However, the lexicon word count

[2]A large amount of options is available for visualizations, but this will not be expanded upon in this thesis. However, one very popular option is the word cloud approach which have been used on the frontpage with the 100 most used words in the NYT2 dataset.

[3]Once again this is only a quick run through of the data management process. For full details see the Main.R file in the github repository: https://github.com/holle94/Speciale

[4]http://www.wjh.harvard.edu/ inquirer/homecat.htm

[5]The Harvard General Inquirer would be optimal for comparison, however, it is not a free access software. The used lexicons in this thesis are freely available and contained in the *tidytext*-package in R.

approach determines that they have similar meaning. This also showcases one of the limitations of positive words, because they often are used to describe negative scenarios.

When the desired lexicon is selected the daily sentiment indexes can be created by counting the amount of words belonging to a certain category. The methods for data collection utilized in this project results in large variations of text data on different days. To solve this problem the different sentiments are normalized by the total amount of words on a given day.

The sentiment indices are created using formula 2.1. s_it is the category of words for article/headline/tweet *i* at the time *t*, w is the total amount of words in article *i* at the time *t*. Six of these categories are created for the word categories: positive, negative, uncertain, litigious, constraining and superfluous.

$$S_t = \frac{\sum_i s_{it}}{\sum_i w_{it}} \tag{2.1}$$

Additionally, a sentiment index, which takes into consideration the relationship between positive and negative words, is created. It is shown in Equation 2.2. P_t is sentiment index from positive words at time *t* and N_t is for negative words at time *t*[6]. This is the same approach as used in Garcia (2013). The SI_t index will simply be referred to as the score index.

$$SI_t = P_t - N_t = \frac{\sum_i p_{it}}{\sum_i w_{it}} - \frac{\sum_i n_{it}}{\sum_i w_{it}} \tag{2.2}$$

2.2 Regression model

In order to test if sentiment indices can forecast the daily returns of the Dow Jones Industrial Average *(Dow)* vector autoregression models (VAR) are used. The VAR models estimated to determine the effect of the sentiment indices are given by equation 2.3.

$$\mathbf{Dow_t} = v + \Phi(L)\mathbf{Y_t} + \lambda\mathbf{X_{t-1}} + \epsilon_\mathbf{t} \tag{2.3}$$

Where the endogenous variables are the lagged values of *Dow* and the selected sentiment index: $\mathbf{Y_t}' = [Dow_t\ S_t]$. A lag operator with five lags is defined: $\Phi(L) = (\Phi_1 L + \Phi_2 L^2 + \Phi_3 L^3 + \Phi_4 L^4 + \Phi_5 L^5)$. The vector X_t contains the exogen variables: a day of the week dummy and a measure of volatility. v is an intercept and ϵ_t is the error term. All the VAR models contain five lags of the endogen variables and one lag of the exogen variables[7]

GARCH-effects

The error term ϵ_t in 2.3 is often assumed to be a white noise process. However, financial times series often suffers from volatility clustering. This violates the assumption of homoskedasticity. A non-constant variance of ϵ_t is modelled to solve this problem, this is done by Generalized Autoregressive Conditional Heteroscedastic (GARCH). The error term is modeled as equation 2.4 where v_t is a white noise process and the time-varying variance is given by h_t.

$$\epsilon_t = v_t\sqrt{h_t} \tag{2.4}$$

[6] All the constructed sentiment variables are stationary as indicated by augmented Dickey Fuller test (ADF-test)

[7] Five lags is selected for a number of reasons: To compare to articles such as Tetlock (2007), to capture effects from the preceding trading week and due to selection criteria such as AIC. SBC select a more parsimonious model with only one lag, using this have a negative impact on the forecasting compared to the highly parametrized VAR models.

h_t is modeled as an ARMA(p,q) process, given by equation 2.5.

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i} \tag{2.5}$$

2.3 Dimensionality Reduction

It is complicated to identify which sentiment indices are relevant and which are redundant for forecasting the daily stock price returns. Including all the variables in the model estimation will lead to an overparameterized model, resulting in additional estimation errors. This section examines methods to solve the problem using dimensionality reduction. Additionally, these processes are used to detect underlying unobservable factors, the following methods seeks to create a linear combination of the sentiment scores.

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised learning algorithm, meaning for every observation there is different units of measurement x_t but no response variable y_t. PCA performs dimension reduction without having any response variable. The independent variables are converted into a smaller number of principal components (PC). Having a number of different variables $X_1, X_2, \ldots, X_p$ the first principal component is given by 2.6[8]. The component is a linear combination of the different variables which explains the largest part of the variance. The second principal component can be found as the linear combination of $X_1, X_2, \ldots, X_p$ which has the highest variance and is uncorrelated with the first PC. The second PC is directional orthogonal on the first PC.

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p \tag{2.6}$$

$\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$ are the loadings of the first principal component. The linear combination is normalized by having the restriction $\sum_{j=1}^{p} \phi_{j,1}^2 = 1$. This is done since arbitrarily large values of the loading could lead to an arbitrarily large variance of the component(James et al., 2014, p.375-378). Given all this PCA solves:

$$\max_{\phi_{11}, \phi_{21}, \ldots, \phi_{p1}} Var(Z_t) \; Subject \; to \; \sum_{j=1}^{p} \phi_{j,1}^2 = 1 \tag{2.7}$$

The PC is also referred to as scores and indicates the coordinates of the new components.

This approach enables the construction of a single variable (PC1) which can be used for estimation and forecasting. Using this the component incorporates information from all the sentiment scores without having to directly estimate all the values in the VAR models.

2.3.2 Partial Least Squares

A drawback of the Principal Component Analysis is that the computed components does not take the response variable into consideration. There is no guarantee that the PC explaining the highest variance in the predictors have any predictive power. An alternative approach to the PCA is the partial least squares regression (PLS).

[8] In matrix form : $T = XP$, easier for comparison of the PLS.

PLS shares some similarities with PCA as it compresses the information stored in the dependent variables. However, PLS aims to produce variables that captures the most variance in both the independent variables and dependent variables, while maximizing the covariance between them. The primary focus will be on extracting the scores. This is done by extracting the weights w from the cross product of the matrix X and Y:

$$S = X^T Y \tag{2.8}$$

The projection of the X matrix on these weights yields the X scores. The scores t serve the same purpose of as the scores in PCA. w describes the directions with most variation in the cross product between X and Y. The scores t are again the coordinates of the new components.

$$t = Xw \tag{2.9}$$

Additionally, the loadings can be obtained by regressing X and Y against the score $p = X^T t$ and $q = Y^T t$. It should be noted that components in PLS are extracted sequentially, whereas the desired amount of PC's in PCA can be calculated in one step. The product of the scores and loadings is subtracted from the X and Y matrices. This deflation removes the information related to the first component and the second component can now be calculated (Wehrens, 2011, p.155-158). The deflation of X is shown in equation 2.10.

$$E_{n-1} = X - tp^T \tag{2.10}$$

2.3.3 Penalized regressions

Estimating the full VAR models may lead to poor forecasting performance due to parameter estimation uncertainty, even with reduced dimensions. To reduce the problem of uncertainty penalized regressions can be used. These methods reduce the variance of the forecast by shrinking parameters towards zero. Elastic net regression will be used in this project. It is a combination of ridge and lasso regression. Elastic net enables both variable selection from the lasso and the ability to shrink correlated predictors with the ridge regression. The parameters in elastic net regression are estimated by minimizing equation 2.11. $\alpha|\Phi_j, \lambda|$ is the lasso part and $(1-\alpha)[\Phi_j, \lambda]^2$ is the ridge part. Lasso can do variable selection by shrinking the parameters to zero due to taking the magnitudes of coefficients into account.

$$RSS + \lambda \sum_{j=1}^{P} (\alpha|\Phi_j, \lambda| + (1-\alpha)[\Phi_j, \lambda]^2) \tag{2.11}$$

$\lambda \geq 0$ is a tuning parameter. When the parameter is 0 there is no penalty and standard least squares will be estimated. However, when λ grows so does the penalty and the coefficients estimated by elastic net will approach zero. α indicates the trade-off between ridge and lasso. γ and α are determined separately through cross-validation (Silge and Robinson, 2019, p.214-223)[9].

2.4 Evaluation criteria

The following section will present the used methods to evaluate the performance of the regression models both in-sample and out-of-sample.

[9] cross-validation is done in a rolling one step ahead forecast, where the test data is divided into three phases. The first for estimation, the second for determining the penalty γ and α and the last for evaluating the forecast.

2.4.1 Granger Causality

Granger Causality is used in order to determine if a time series variable x_t is usable in forecasting another time series y_t. The test is conducted by estimating a model containing lags of both y_t and x_t, shown by equation 2.12.

$$\mathbf{y_t} = v + \beta(L)\mathbf{y_t} + \gamma(L)\mathbf{x_t} + \epsilon_\mathbf{t} \tag{2.12}$$

In the equation with a lag length of *p*, x_t does not Granger cause y_t if $\gamma(L) = 0$. With stationary variables a standard F-test i used to test: $\gamma(L) = (\gamma_1 = \gamma_2 = \ldots = \gamma_p) = 0$. If any of these values are statistically different from zero, the variable Granger causes and can potentially be used for improving forecasting (Enders, 2015, p.305-306). It should be noted that Granger Causality only takes into consideration linear relationships between variables.

2.4.2 Clark-West test

In order to evaluate the performance of the models containing information from the sentiment indices *mean squared prediction error* (MSPE) will be calculated. This will be done for both the VAR model and a benchmark model after which a ratio will be calculated. If equation 2.13 is less than 1, then the proposed VAR model have lower MSPE than the benchmark model.

$$Ratio = \frac{MSPE_{VarModel}}{MSPE_{Benchmark}} \tag{2.13}$$

To determine if the proposed model have statistically superior ability to forecast the *Dow* compared to a benchmark model, the Diebold-Mariano test(DM) can be used. The loss from period i from model 1 compared to model 2 is $d_i = g(e_{1i}^2(j)) - g(e_{2i}^2(j))$. The mean loss function is given by:

$$\bar{d} = \frac{1}{H}\sum_{i=1}^{H}[g(e_{1i}^2(j)) - g(e_{2i}^2(j))] \tag{2.14}$$

The null hypothesis(H_0) in the DM test is that of equal forecast accuracy of the two models: $\bar{d} = 0$. The alternative hypothesis (H_1) is a one-sided test which states that the proposed model have superior forecast performance compared to the benchmark model. The Diebold Mariano statistic for one-step-ahead forecast is given by equation 2.15 .

$$DM = \frac{\bar{d}}{\sqrt{\frac{\gamma_0 + 2\gamma_1 + \ldots + 2\gamma_q}{H-1}}} \tag{2.15}$$

The benchmark model for comparison will be a simple AR(1)-model of the *Dow*, resulting in nested models. Under the null hypothesis the models should predict MSPE equally. The larger model will however contain extra uncertain due to the additional non-relevant parameters. The modified DM-test proposed by Clark and West (2007) will instead be used to adjust for the uncertainty. The modified test is conducted by subtracting the difference between the forecast (f_1 and f_2) from errors of the non-nested model. This series is shown in equation 2.16.

$$z_i = (e_{1i})^2 - [(e_{1i})^2 - (f_{1i} - f_{1i})^2] \;\; i = 1, \ldots, H \tag{2.16}$$

Like in the standard DM-test the null hypothesis is equal forecast performance, resulting in z_i beaning 0. The alternative hypothesis is that the non-nested model has superior performance (Enders, 2015, p.86-87).

Mean Directional Accuracy

In this project the ability of the estimated models to forecast directional changes (up or down movements) in the *Dow* is evaluated by calculating the *mean directional accuracy* (MDA). The MDA is given by equation 2.17

$$\frac{1}{H}\sum_{t}\mathbf{1}_{sign(A_t-A_{t-1})==sign(F_t-A_{t-1})} \tag{2.17}$$

A_t is the actual value of the *Dow* and F_t is the forecasted value at time *t*. *sign()* is a signum function and 1 is an indicator function. If the forecasted directional sign is identically to the actual direction the function returns the value 1. This is done for periods t to $t+H$ and averaged resulting in an MDA between 0 and 1 .

2.4.3 Pesaran-Timmermann test

The Pesaran-Timmermann (PT) test is used in order to determine if the proposed VAR models are statistically able to forecast the direction of the *Dow*. Pesaran and Timmermann (1992) proposed a non-parametric test for this purpose. The PT test is given by equation 2.18, for a time series y_t and the forecasted values of the series x_t.

$$S_n = \frac{\hat{P}-\hat{P}_*}{[v\hat{a}r(\hat{P})-v\hat{a}r(\hat{P}_*)]^{0,5}} \tag{2.18}$$

where

$$\hat{P} = n^{-1}\sum_{t=1}^{n} I(y_t x_t)$$

$$\hat{P}_* = \hat{P}_y\hat{P}_x + (1-\hat{P}_y)(1-\hat{P}_x)$$

$$\hat{P}_y = n^{-1}\sum_{t=1}^{n} I(y_t) \; and \; \hat{P}_x = n^{-1}\sum_{t=1}^{n} I(x_t)$$

$$v\hat{a}r(\hat{P}) = n^{-1}\hat{P}_*(1-\hat{P}_*)$$

$$v\hat{a}r(\hat{P}*) = n^{-1}(2\hat{P}_y-1)^2\hat{P}_x(1-\hat{P}_x) + n^{-1}(2\hat{P}_x-1)^2\hat{P}_y(1-\hat{P}_y) + 4n^{-2}\hat{P}_y\hat{P}_x + (1-\hat{P}_y)(1-\hat{P}_x)$$

$$I(\cdot) = \begin{cases} 1 & \text{if } \cdot > 0 \\ 0 & \text{otherwise} \end{cases}$$

In the above equations $\hat{P}$ is the proportion of times that the direction of y_t is forecasted correctly. $\hat{P}_y$ and $\hat{P}_x$ are the probability that y_t and x_t are larger than 0: $P_y = Pr(y_t > 0)$, $P_x = Pr(x_t > 0)$). $\hat{P}_*$ is the probability that the sign is forecasted correctly. This is calculated as the probability that both y_t and x_t are above 1 added the probability that they are not. The null hypothesis for the PT test is that S_h follows a normal distribution. The model x_t is not able to forecast the directional change of the series y_t. If the null hypothesis is rejected on a α- significance level, the alternative hypothesis states that the model is able to forecast the direction (Pesaran and Timmermann, 1992, p.461-463).

Chapter 3

Data

3.1 Text data

This section contains descriptions of the text data used to create the sentiment indices such as time frame, data source and how the data was collected[1].

3.1.1 New York Times

Three financial text datasets in this project was collected using the freely available *The New York Times* API[2]. The API enables the option to do article searches in specific time intervals and on specific search words. The primary focus in this project will be on the period 01/01/2011-31/12/2019. Furthermore the API is used to select specific sections from *The New York Times* in which searches will be conducted. It should be noted that this method only downloads the headline and the byline and not the entire article. Tetlock (2007) only uses a specific column from *The Wall Street Journal*. However, due to the accessibility this project instead tries to capture the financial mood by aggregating a large amount of articles. Another drawback is that this method might catch articles with little financial relevance. This is assumed to only account for a small fraction since the selected sections primarily focuses on financial/economic news and search words. This process has some limitations: The API have a hourly call limit which significantly increases the download time.

In the first dataset (NYT1) all the headlines and bylines in the business section, in the period 2011-2019, are downloaded. For an extract of the data see Appendix: A.1. This yields approximately 97,000 headlines and bylines[3]. The second dataset (NYT2) is collected from the same time interval and section. However, this time the information from all articles containing the words stock and market are downloaded. The search in the period 2011-2019 yields approximately 51,000 headlines and bylines. The third dataset (NYT3) searches only for articles containing the word *economy* in the same time period. In the NYT3 dataset the section has not been specified, since *The New York Times* contains many other sections of interest, especially

[1] All data and codes are available on https://github.com/holle94/Speciale. It should be noted that data from Thomson Reuters Eikon is not available there, instead data from Yahoo finance is supplied.

[2] API and guides available on: https://developer.nytimes.com/

[3] The original count was 107,000, but due to the search pattern of the API some duplicates appear, which were removed. This was the case for all three datasets.

when related to the economy. This set contains approximately 63,500 headlines and bylines. The following tables shows a quick representation of the text data process from the NYT2 dataset.

Table 3.1 Most used words in headlines and bylines from NYT

Word	market	company	billion	deal	business
Count	4919	4093	3882	3544	3401

Note: These are the most occurring words once stop words etc. have been removed.

Table 3.1 displays the five most used words once stop words are removed. They all appear to be of some financial or economic character. It is therefore assumed that the majority of the articles are of some relevance. However, these are neutral words and does not provide much information regarding the financial media sentiment. To further narrow down the relevant words the lexicons approach from the previous chapter is used. Table 3.2 presents the most used sentiment words with the *Loughran and McDonald* lexicon.

Table 3.2 Most used words with the Loughran Lexicon

Word	may	could	crisis	against	good
Count	2872	2560	1171	1059	786

Note: These are the most occurring sentiment words contained in the NYT2 dataset.

Using the *Loughran and McDonald* lexicon approach divides the remaining words into six different categories shown in table 3.3.

Table 3.3 Distribution of words using the Loughran Lexicon

constraining	litigious	negative	positive	superfluous	uncertainty
2040	8849	39525	10431	35	5368

Note: distribution of words in the different sentiment indices for the NYT2 dataset.

The news headlines and bylines are dominated by negative words. However, they also contain a lot of positive, litigious and to a minor degree uncertain words. Litigious words are primarily related to legal actions of specific companies and might not relate to the broader stock market. However, the ability of the different categories to forecast stock market movements will be examined in the next chapter. Superfluous words are assumed to have little impact, since they only appear 35 times while the dataset contains 2263 daily observations of the *Dow*. However, they will still be included since they might add information when combined linearly with other variables in the PCA and PLS.

Once this process is done, the amount of words in each category is summarized for each day and normalized by the total amount of words on the same day[4]. An example of the negative sentiment index can be seen in Appendix A.2.

[4]All the constructed sentiment indices are stationary

3.1.2 Reddit

Reddit is an online news aggregator and online discussion forum[5]. Reddit has approximately 430 million active users and contains an extremely large amount of subforums. Reddit data such as comments and posts are available from Google Big Query [6]. Reddit contains additional noise compared to the text data collected from the *The New York Times* API. Posts and comments can be made by anyone and there is no guarantee that the statement and information is true[7]. To solve this aggregation of a large amount of posts/comments could once again be a solution to capture the sentiment. The subforum search is also important, since searching for specific words across all subforums might lead to a lot of unrelated text data. Another option is to only download the highest rated posts/comments, since Reddit has a rating system. Two datasets were considered for this project. The first was downloaded from Big Query containing posts in the subforum Investing. This subforum contains posts about investing and general financial news. The second was downloaded from Kaggle.com containing the daily top 25 highest rated news from the Reddit World News Channel (RWNC). This sub forum collects major news from around the world. Only the results for the second dataset are presented, they were better performing. The RWNC dataset stretches from 09/06/2008-01/07/2016 and contains 73600 news headlines. The RWNC have 24 million followers, whereas the investing forum have 1 million.

3.1.3 Twitter

Twitter is a social media platform where users post and interact with their network through tweets. Twitter have more than 320 million monthly active users[8]. These includes active users such as investors, economists and politicians. Most prominent lately is the current President of the united states, Donald Trump.

The Twitter data used in this project is the same dataset used in Mudinas et al. (2018) [9]. This dataset covers the time period 06/05/2014 to 01/02/2015 and contains tweets with cashtag ($) and a stock ticker from stocks in the DJIA. The dataset covers a rather short time period. This is primarily due to the access to historic tweets being rather restricted. A basic twitter developer account only has access to tweets from the last 30 days. Students can apply for premium accounts and gain access to historical tweets stretching back more than 30 days [10]. However, this only allows for 5000 tweet downloads a month, which in the case of this project have been used to impute dates in the dataset with no tweet observations[11]. To receive a larger amount of tweet data an enterprise account is required, which requires monthly payments depending on the number of downloaded tweets. Despite the short timeframe the dataset includes more than 1.2 million tweets. Tweets have a text limit of 140 characters limiting the amount of text information. Once again, this problem is hopefully solved by the aggregation of large amounts of tweets.

[5] https://www.redditinc.com/

[6] This is not at free service. However, Google awards 500$ worth of credit to new users, which if managed properly is more than enough to download a decent seize dataset.

[7] Additionally the forum often contains a large amount of bots and deleted comments.

[8] https://about.twitter.com/

[9] https://github.com/AndMu/Market-Wisdom

[10] https://developer.twitter.com/en/products/tweets

[11] Due to the cashtag ($) search only being available to enterprise version, the search was simply done on the tickers: AAPL and GOOG. This resulted in some unrelated tweets, which had to be filtered out. This resulted in imputations of only around 2000 tweets.

Due to the nature of the downloaded tweets, the effect of sentiment can be evaluated on the daily return of individual stocks in DJIA. In this project the stocks Apple and Google will be examined. It would be of great interest to examine more individual stocks; this is not done since the monthly cap on tweet downloads have primarily been used to do imputation on missing values for Google and Apple.

3.1.4 Datasets

When comparing the general purpose *bing* lexicon and the financial focused *Loughran and McDonald* lexicon some clear differences appear. The *bing* lexicon classifies a larger proportion of the words to the negative and positive categories. For the NYT2 dataset the mean of the daily percentage of negative words is 5.82% and 4.31% for positive. When using the *Loughran and McDonald* these values are only 4.48% and 1.21%[12]. For a comparison of sample statistics for the sentiment indices see Appendix: table A.1. This supports the findings in Loughran and McDonald (2011) that negative words are often misclassified in financial context. The positive word list created by Loughran and McDonald (2011) appear to be more picky due to the large difference in the proprtion compared to *bing*. It should be noted that in general both lexicons find a higher average number of negative words than positive words for the different datasets. Seven datasets will be used in total to evaluate the effect of media sentiment on the *Dow*. The datasets are shown in table 3.4 where they are collected from, time period and what kind of text data they contain. The selected time periods are not optimal for comparison, since they contain varying time intervals though with some time overlaps. However, as mentioned earlier the accessibility of the text data is often rather restricted.

Table 3.4 Datasets used for evaluating the effect of sentiment on stock price movement

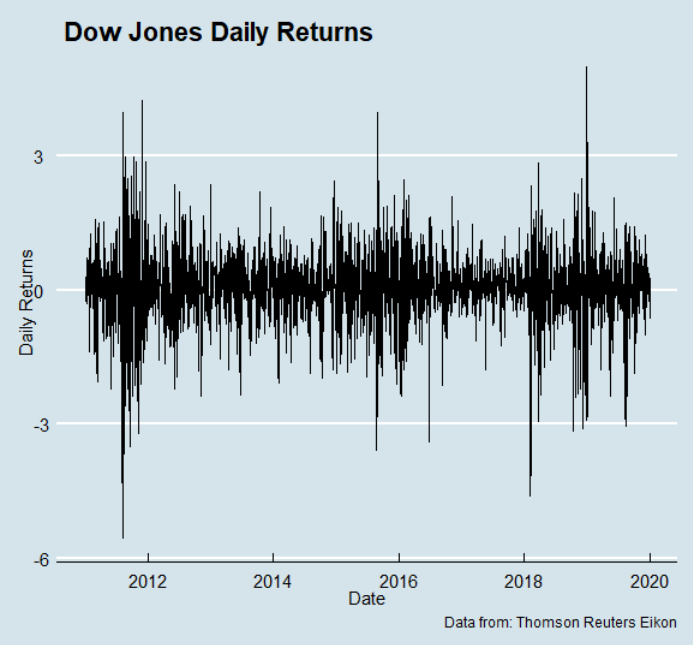
Dataset	Time frame	Text data
FT Full	01.01.11-31.12.19	Headlines and Byline
FT Mix	01.01.11-31.12.19	Headlines and Byline
FT Econ	01.01.11-31.12.19	Headlines and Byline
RWNC	09.06.08-01.07.16	Headlines
Dow Jones Twitter	06.05.14-01.02.15	Full Tweets
Apple Twitter	06.05.14-01.02.15	Full Tweets
Google Twitter	06.05.14-01.02.15	Full Tweets

3.2 Dow Jones Industrial Average

The data for the DJIA returns has been collected from Thomson Reuters Eikon. The daily returns are shown in plot 3.1.

[12]The results are similar for all the other datasets

Figure 3.1



From the plot the time series appears to be stationary, which is confirmed by running an augmented Dickey Fuller test(ADF). However, the series shows periods of high volatility followed by rather tranquil periods. This appears as an indicator that the series have a non-constant variance[13].

The stock markets are closed on weekends and holidays, whereas the sentiment from news and social medias are collected on all days. To incorporate the information from text on the weekends, these are aggregated and added to the preceding Friday. The text data is collected from the entire preceding day. Additionally, a version was examined using text information only available between market close and market open the following day. However, this did not affect the results significantly. Sample statistics of the three different time periods of the *Dow* are available in Appendix: table A.2.

PCA and PLS as described create new latent variables. The first PC and PLS component load primarily on the negative sentiment index with some additional loading on the positive and litigious indices. An example of loadings of the first two components from PCA and PLS for the NYT1 dataset can be found in Appendix: table A.3.

[13] An AR-model is estimated to determine if any heteroskedastic appear in the series. The residuals do not appear to have any significant correlations, which is confirmed by using Ljung-Box test. However, when examining the squared residuals, the Ljung-Box finds significant correlations. To deal with the heteroskedastic a GARCH(1,1) model is estimated. The GARCH is selected since it is more parsimonious than a higher order ARCH model, making it easier to estimate. Auto Correlation plots of the residuals and the squared residuals can be found in Appendix: Figure A.3 and A.4.

Chapter 4

Results

In this chapter the models will be estimated to examine if statistical relationships are present between the daily return of the DJIA (*Dow*) and lags of the media sentiment measure S_t. First tests are made to examine if the different sentiment indices Granger cause the daily changes in the Dow Jones index. Secondly the models are estimated in-sample and the parameters are inspected and compared to the results from Tetlock (2007) and Garcia (2013). Third the models are estimated out-of-sample to eliminate any potential look-ahead bias and to evaluate forecast performance.

4.1 In-sample

The models will be estimated in-sample on the entire length of the time periods. Firstly to examine if any of the variables Granger cause the *Dow* and secondly to estimate the VAR-models to examine the parameters, their significance and if any clear patterns emerges.

4.1.1 Granger Causality

Table 4.1 presents the results of the Granger Causality tests for the different sentiment indices and text data sources. The test was made with five lags of the indices and the *Dow*.

Table 4.1 Granger-Causality of sentiment indices on the *Dow*

Dataset	Score	P	N	U	C	L	S	AFINN	PC	PLS
NYT 1	0.45	0.23	0.76	0.63	0.10	0.79	0.92	0.61	0.91	0.95
NYT 2	0.17	0.58	0.18	0.46	0.53	0.38	***0.015***	0.12	0.11	0.68
NYT 3	0,44	**0,087**	0,29	0,95	0,14	0,55	0,45	0,76	0,75	0.45
RWNC	0,34	0,54	0,48	0,99	0,61	**0,057**	0,95	0,37	0,19	0.12
Dow T	0,86	0,85	0,83	0,74	0,68	0,79	0,17	0.39	0,91	0.16
Google T	***0,039***	0,69	0,12	***0,038***	0,73	0,69	0,13	0,61	0,25	**0.058**
Apple T	0,49	0,54	0,33	0,14	0,27	0,85	0,75	0,45	0,17	0.22

The NYT1 contains all headlines from the Business section of the New York times. The NYT2 is more focused, selecting only headline containing the words: market and stock. NYT3 i similar, selecting only headlines containing the word: Economy. Numbers in bold indicates significance on a 10% significance level. Bold and italic indicates on a 5% significance level. .

Table.4.1 shows that some variables Granger cause *Dow* and indicates that some of the sentiment indices can be used for forecasting future values of the *Dow*. However, no clear pattern appears, different indices Granger cause from different text sources and time frames. Superfluous words Granger cause the *Dow* on a 5%-significance level for the NYT2 dataset. Once again this might not lead to good out-of-sample performance due to the few words in this category. The Granger tests for daily changes in Google and Apple display quite different results, despite both being individual stocks in the same time period. The causality test indicates that PLS is to prefer over the PCA. This might be misleading since the sentiment indices in the PLS are modelled against the *Dow* one period ahead. This look-ahead bias is eliminated in the out-of-sample regression. The next step is to estimate equation 2.3 with the different sentiment indices.

Estimation of *Dow*

Equation 2.3 is estimated in-sample using five lags of the *Dow* and the different sentiment indices[1,2].

$$\mathbf{Dow_t} = v + \Phi(L)\mathbf{Y_t} + \lambda\mathbf{X_{t-1}} + \epsilon_\mathbf{t}$$

The below tables summarize the estimated parameters from the proposed models containing different sentiment indice[3]. This is only some of the estimates, the remainder can be found in the Appendix.

Table 4.2 Parameter estimates with media sentiment from NYT2

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	-0.001	0.975	-0.000	0.99	0.005	0.82	0.028	**0.09**	-0.043	0.25
S_{t-2}	0.012	0.23	0.014	0.19	-0.001	0.95	0.023	0.14	0.013	0.44
S_{t-3}	0.015	0.11	0.016	0.14	-0.005	0.86	0.017	0.27	-0.011	0.51
S_{t-4}	-0.019	**0.05**	-0.019	**0.06**	0.023	0.39	-0.023	0.17	0.086	0.99
S_{t-5}	-0.005	0.62	0.001	0.91	0.036	0.11	-0.004	0.79	-0.021	0.89

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Table 4.2 contains the estimates from sentiment indices created from text in the NYT dataset. Only a few of the parameters are significant on a 10% level. The standard dictionary approach indices, such as negative and positive, seems to primarily capture some information in lags $t-4$. The indices such as PC and PLS instead capture some information in lag $t-1$.

[1]The second components PC2 and PLS2 were also included in the VAR model, this did not improve performance of the models.

[2]The characteristic roots are computed for all the VAR models to ensure they. All the used models are found to be stationary.

[3]Only the negative and positive indices are included, since the remainder performed poorly.

Table 4.3 Parameter estimates with media sentiment from RWNC

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	-0.001	0.97	0.001	0.91	0.025	0.71	0.024	0.26	0.033	**0.07**
S_{t-2}	-0.013	0.44	-0.013	0.48	0.025	0.72	0.017	0.53	0.033	0.74
S_{t-3}	-0.011	0.50	-0.007	0.68	0.077	0.20	-0.011	0.92	0.036	0.63
S_{t-4}	0.004	0.82	0.007	0.70	0.046	0.49	0.011	0.66	0.033	0.13
S_{t-5}	0.035	***0.04***	0.032	**0.06**	-0.077	0.19	0.055	**0.05**	0.029	0.22

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

The results from Table 4.3 are quite like the results above. However, this time the information seems to primarily be found for lag $t-5$, but still some information is found from lag $t-1$ with the dimensionality reduction methods.

Table 4.4 Parameter estimates with media sentiment from tweets related to Google

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	0.171	0.327	0.130	0.40	-0.143	0.65	0.103	0.30	-0.432	**0.10**
S_{t-2}	-0.365	0.016	-0.300	***0.02***	0.141	0.71	0.038	0.67	0.054	0.78
S_{t-3}	0.043	0.778	-0.063	0.65	-0.230	0.29	0.023	0.81	0.405	**0.09**
S_{t-4}	0.280	0.102	0.191	0.23	-0.078	0.78	-0.086	0.35	0.404	***0.02***
S_{t-5}	0.197	**0.068**	0.194	**0.08**	0.338	0.18	-0.092	0.24	-0.332	0.15

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Summarized the above tables support the conclusion that the sentiment indices due contain some information in regards to future returns of the *Dow*. Once again, no clear pattern appears. The significant coefficients are not uniform across the different indices. Despite being estimated against the future returns of the *Dow*, the PLS does not seem to have superior performance. In general, the significant parameters appear to be mostly in lag $t-4$ and $t-5$ with a few at lag $t-1$ for PC and PLS. This is in contrast to the results found by Tetlock (2007) and Garcia (2013). They find a strong statistically significant initial effect from $t-1$. However, they also find a reversal of the initial effect over the following 4 days with significant parameters at lag $t-4$ and for some indices at lag $t-5$.

4.2 Out-of-sample

The in-sample regression indicates that some relation exists between some of the sentiment indices and future values of *Dow*. In this section the sentiment indices will be used to forecast out-of-sample. The out-of-sample is based on an expanding estimation window. To prevent any look-ahead bias the out-of-sample is estimated recursively and using only the information that is available at the time of the forecast. The forecast will be done for one step ahead(h=1). Given that the PLS component is modelled with both variation in the dependent and independent variables. This will result in potential look-ahead bias. To solve this problem the score from PLS

is estimated using only information available at the time of forecast. Thus, the weight matrix *w* from the preceding periods is used, since the weight cannot be computed due to the value of *Dow* one step ahead being unknown at the time of forecast.

The estimated models will be compared to an AR(1)-model of the *Dow* as a benchmark model. This comparison will be made since parsimonious models often produces better forecasts than heavily parameterized models [4].

The used datasets have different estimation and evaluation periods due to the difference in the length of the datasets. For the datasets with text data from *The New York Times*, the initial 1163 trading days are used for estimating the model and the remaining 1000 days are used for evaluating the forecast performance. The reddit dataset uses the first 1031 trading days for estimation and the remaining 1000 for evaluation. For the datasets from Twitter the first 100 days are used for estimating the model and the remaining 86 days are used for evaluating.

Table 4.5

	Score		N/P		PC		PLS		AR(1)	
	Acc.	MSPE	Acc.	MSPE	Acc.	MSPE	Acc.	MSPE	Acc.	MSPE
NYT1	0.51	1.02	0.52	1.02	0.52	1.02	0.52	1.00	***0.54***	0.69
NYT2	0.51	1.01	0.52	1.02	0.52	1.01	**0.53**	0.99	***0.54***	0.69
NYT3	0.51	1.02	**0.53**	1.01	0.51	1.02	0.53	1.01	***0.54***	0.69
RWNC	0.50	1.02	0.51	1.01	0.515	1.01	**0.53**	1.00	0.51	0.66
Dow T	0.52	1.12	0.48	1.12	0.53	1.13	0.53	1.10	0.49	0.91
Apple T	0.53	1.06	0.52	1.07	0.51	1.08	0.49	1.06	0.45	2.67
Google T	***0.60***	1.01	0.51	1.04(u)	0.48	1.07	0.53	1.03	0.48	2.03

Bold accuracy scores are Pesaran-Timmermann test significanse on 10%-level. Bold and italic accuracy scores are significanse on a 5% confidence level. Bold MSPE ratios are Clark-Vest test significanse on 10%. Bold and italic accuracy scores are on a 5% confidence level. The MSPE of AR(1) is not a ratio, but the true MSPE value of the model. The N/P is the best performing of the negative or positive indices for each dataset.

Table 4.5 presents the results of the out-of-sample regression. Acc. is the directional accuracy and the significance is tested using the PT-test. MSPE is a ratio between the MSPE of the models containing sentiment and the AR(1)-model. A ratio above 1 indicates lower MSPE of the AR-model and vice versa. The significance of superior forecast performance is tested using the Clark-West test. The table illustrates that the news indices contain some information for forecasting. In general they perform poorly compared to the simple model. This is especially clear when examining the first three datasets. For the NYT2 dataset, the PLS achieves a statistically significant mean directional accuracy. However, this accuracy is lower than the one obtained by the simple AR(1)-model. The same is apparent from the positive index for the NYT2 dataset. None of the models have a lower MSPE than the AR(1)-model. This is likely due to the large amount of estimated parameters.

For the RWNC the PLS is able to achieve a significant accuracy at 53%, which is better than that of the AR(1)-model. However, it has approximately the same MSPE resulting in an MSPE ratio of 1.

As mentioned earlier the Twitter data enables the option to examine not only daily changes in the *Dow*, but also the daily changes in stocks of Google and Apple. For the *Dow* the AR(1)-model

[4]The AR(1)-benchmark model also includes day of the week dummy and GARCH(1,1) estimates.

performs poorly with an acc. of 49% and is outperformed by both the score index and the PC and PLS indices. However, these accuracies are not significant from the PT-test due to the short sample size. Additionally, the MSPE of the models containing the lagged sentiment is quite a lot higher than for the AR(1)-model. The models for predicting the daily change of Apple performs similar, none of the accuracies or ratios are significant. The MSPE ratios are all above 1. The results for Google in general are similar. However, using lagged values of the sentiment score an accuracy of 60 % is achieved, which is significant at a 5%-level. Once again it should be kept in mind that the Twitter dataset contains few time observations reducing the statistical power.

Categories with few words, such as the superfluous category, seems to Granger cause the *Dow* at times. However, they perform badly in OOS-prediction[5]. The next section will aim to improve on the out-of-sample forecasts by using penalized regression.

4.2.1 Penalized regression

Using elastic net regression all the sentiment variables are included in the forecast of *Dow*. Normally, this would have led to an extremely overparameterized model. Due to the models ability to do variable selection this will be less of an issue. Using this method eliminates the issue of determining what variables to include. Furthermore, it allows for combinations of lags from different indices[6].

Figure A.5 in the Appendix illustrates how the different parameters are penalized in the VAR model for the NYT3 dataset using all the sentiment indices. The figure represents the lags of the *Dow*, negative, positive, litigious, uncertain, constraining and superfluous. The intensity of the colored squares indicates the penalty. White squares indicates parameters shrunken to zero though it is hard to determine from the plot if parameters are simply heavily penalized or set to zero. Figure A.6 in the Appendix displays the shrunken parameters.

The results of the penalized regression is shown in table 4.6.

Table 4.6 Penalized regression

	Full model	
	Acc.	MSPE
NYT1	0.54	1.00
NYT2	0.54	0.99
NYT3	0.53	1.00
RWNC	0.525	0.99
Dow Twitter	0.46	0.98
Google Twitter	0.47	1.04
Apple Twitter	0.47	0.99

Bold accuracy scores are Pesaran-Timmermann test significanse on 10% confidence level. Bold and italic accuracy scores are significanse on a 5%. Bold MSPE ratios are Clark-Vest test significanse on 10%. Bold and italic accuracy scores are on a 5% confidence level.

As seen from the table none of the accuracies are significant. This is either due to low acc. or in the case of the NYT2 dataset, which primarily predicts positive daily returns. The penalized

[5] These results are not included since they performed considerably worse than the other options.

[6] An alternative approach would have been to just use lags of the PC1 and PLS1 in the penalized regression, the runtime of these models are quite long. This was not done due to time constraints

regression from NYT1 shrinks all parameters to zero which results in a rolling mean model of the *Dow*. This results in all forecasts being above 0, which is the cause of a non-significant accuracy, despite being above 54%. Using penalized regression also have a negative impact on the short Twitter datasets, which leads to a significant decrease in accuracy. The penalized regression approach can reduce the MSPE, though none of them are significantly better than the AR(1)-model.

Chapter 5

Discussion

This section is dedicated to discussing the obtained results and some potential reasons why they differ from those of other papers, such as Tetlock (2007) and Garcia (2013). Furthermore, additional methods and what the future holds for this topic will be discussed.

One reason of this might be, as mentioned earlier, the use of headlines and short snippets of text. The short text pieces might not be able to capture the mood of the news stories and social media posts. Headlines might be misleading lexicon wise, whereas an entire column might summaries the senitment better. A headline might be very ironic or use positive words to describe a negative situation, with the use of such things throughout an entire section being unlikely. A distinction should be made between the different text data sources. News from large publishers are often more formal and uniform. They undergo thorough proofreading before publication. On the other hand, social medias contain more sarcasm/irony, grammatic errors etc. Misspelled words in posts might be important due to not counting in sentiment scores, since they do not appear in the lexicons. Additionally, when using free news/social media it seems to be important to separate noise from the data. It is of great importance where the data is collected from, such as medias, sections and search words. Searching on specific words through an entire forum or in tweets might lead to poor results. An example of this could be "Open the amazon link in an incognito window to not give reddit referral revenue" with a search for mentions of Amazon on Reddit through Google BigQuery. This tells little of the users feeling of Amazon or their financial performance. Instead the search should be narrowed to specific sections or search using the stock ticker: AMZN. The relationship between the news and daily returns is complex, being very specific about section and search patterns might not always be enough. When including the news from the *DealBook* section in the NYT2 dataset, no variables Granger caused and the forecasting performance deteriorated. This was despite this specific section being focused on financial, economic and political news[1].

Another factor to be considered about the use of sentiment extracted from news is the availability and the frequency. The time periods examined by Tetlock (2007) and Garcia (2013) are 1984-1999 and 1905-2005 respectively. Especially, Garcia (2013) is primarily during periods where news were very uniform and the frequency of news was rather low. Both were prior to the

[1]Many headlines and bylines in this section only contained the sentence "Dealbook: market summary". This however could not be the reason for the poor performance, since identical entries were removed.

creation of Twitter and Reddit, as well as periods where online news media were not available or not very prominent. As noted in Garcia (2013) the effect of news sentiment are strongest during periods of recession. However, the datasets used in this project are all primarily during economic expansion.
The *New York Times* API would have been the easiest option to obtain data from other time periods similar to those examined by Tetlock (2007) and Garcia (2013). To determine if there have been a change in the information contained in financial news or if financial headlines simply does not contain enough information[2]. The availability of news is also of great interest. The use of daily news might simply be of a to low frequency. News and social medias are available all throughout the day which allows for a more instantaneous reaction.

Going forward Twitter appears to be the most promising of the used text data sources. Both due to classic news sources being less uniform and the huge amount of available hourly tweets. Bollen et al. (2013) and Mudinas et al. (2018) achieve improvements to their baseline models using large amounts of tweets. The popularity of Twitter has given rise to other similar services, such as StockTwits. This service is very similar to Twitter but targeted at investors and traders.

The used VAR models could have been extended by including measures of the volume of news regarding certain stocks. The number of daily tweets or other measures of trends could be used. Days with high activity could be an indicator of events with impact on daily return. Tools such as Google Trends could be used to examine this like in Huang et al. (2019). This is a freely available service. However at the time of writing it does only contain observations on a weekly frequency.

Another avenue for further exploration would be to use more complex methods than the lexicon approach. Advanced natural language processing (NLP) models could be implemented to better capture the rich nature of text from news and social media posts. Some of the NLP approaches do not only take each word into consideration, but also the placement of the word in the sentences and the surrounding words. This could possibly be a solution to the headline approach since the methods would be better able to capture irony and hidden meanings.

The choice of models should also be taken into consideration. The Granger Causality test as mentioned only captures linear relationships. Papers such as Bollen et al. (2013) and Mudinas et al. (2018) have had some success using more complex non-linear models. Bollen et al. (2013) use linear combinations of multiple sentiment variables. This yields inferior results compared to only incorporating the best performing index. However, when using combinations in non-linear models this results in improved performance, which leads to the conclusion that non-linear relationships exist between some sentiments.

[2]This was not done due to the time constraints on the NYT API.

Chapter 6

Conclusion

The goal of this thesis was to examine if sentiment extracted from news, forums and social media can be used to forecast the daily return of the *Dow Jones Industrial Average*. The results indicate that some information is to be found in the sentiment extracted from the text data. However, no clear patterns are found. Different sentiment indices Granger cause the *Dow* and are not uniform across different datasets. No clear in-sample pattern was found like those in Tetlock (2007).

The out-of-sample performance showed that including sentiment in the model, a few indices were able to statistically achieve a directional accuracy outperforming the benchmark model. Once again, no clear pattern was found. Different indices from different text datasets improved performance. However, in general the models were not able to outperform the parsimonious benchmark model. Using principal component analysis and partial least squares did not improve the performance of the models, despite being able to estimate underlying relationships between the different sentiment indices. The use of penalized regression to circumvent the problem of overparameterized models had a negative impact on the performance. The sentiment lags containing information were often penalized to hard or set to zero.

The use of media sentiment is an area of great interest and will most definitely receive even more attention in the future. More complex methods might be needed to capture the complex relationship between news and daily returns. The VAR models are unable to capture non-linear relations and might be less desirable compared to deep-learning methods. The lexicon approach might also be a bit outdated due to the increasing rise of social media use. A more advanced approach is needed in order to capture the deep complex meanings of texts.

Bibliography

Bollen, J., Mao, H. and Zeng, X. (2013), 'Twitter mood predicts the stock market', *Journal of Computational Science* 2.

Clark, T. and West, K. (2007), 'Approximately normal tests for equal predictive accuracy in nested models', *Journal of Econometrics* 138(1), 291–311.
URL: *https://EconPapers.repec.org/RePEc:eee:econom:v:138:y:2007:i:1:p:291-311*

Enders, W. (2015), *Applied Econometric Time Series*.

Garcia, D. (2013), 'Sentiment during recessions', *Journal of Finance* 68(3).

Goyal, A. and Welch, I. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *The Review of Financial Studies* 21(4), 1456–1508.

Huang, Y. M., Rojas, R. R. and Convery, D. P. (2019), 'Forecasting stock market movements using google trend searches', *Empirical Economics* .

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014), *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated.

Loughran, T. and McDonald, B. (2011), 'When is a liability not a liability? textual analysis, dictionaries, and 10-ks', *Journal of Finance* 66(1).

Mudinas, A., Zhang, D. and Levene, M. (2018), 'Market trend prediction using sentiment analysis: Lessons learned and paths forward'.

Pesaran, M. H. and Timmermann, A. (1992), 'A simple nonparametric test of predictive performance', *Journal of Business and Economic Statistics* 10, No. 4, 461–465.

Silge, J. and Robinson, D. (2019), *Text Mining with R*.

Tetlock, P. C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *Journal of Finance* 62(3), 1139–1168.

Wehrens, R. (2011), *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*, Springer, Heidelberg.

Appendix A

Appendix

Figure A.1 A snippet of text data from the NYT1 dataset

created_time	snippet	headline	text
2011-01-01	Machines have largely taken over stock market trading, crea...	The New Speed of Money, Reshaping Markets	The New Speed of Money, Reshaping Markets Machines ha...
2011-01-01	The founder of Vanguard says that trying to time the market...	Market Wisdom Applies to E.T.F.'s, Too	Market Wisdom Applies to E.T.F.'s, Too The founder of Vang...
2011-01-01	Experts warn of volatility in state finances and the broader s...	Europe's Young Grow Agitated Over Future Prospects	Europe's Young Grow Agitated Over Future Prospects Expert...
2011-01-01	Unlike ordinary homeowners, developers who suffer costly l...	Real Estate Developers Prosper Despite Defaults	Real Estate Developers Prosper Despite Defaults Unlike ordi...
2011-01-02	Sales campaigns for commercial jets on the global market o...	Diplomats Help Push Sales of Jetliners on the Global Market	Diplomats Help Push Sales of Jetliners on the Global Market...
2011-01-03	Fiat Industrial and the Fiat auto business both gained on M...	Fiat Shares Rise After Company Splits	Fiat Shares Rise After Company Splits Fiat Industrial and the ...
2011-01-03	In a move that might spur greater integration with Chrysler, ...	Fiat Shares Rise After Company Splits	Fiat Shares Rise After Company Splits In a move that might ...
2011-01-03	Apple's sales were up 50 percent in its financial year ended i...	Apple Will Shine in 2011, if Not as Blindingly	Apple Will Shine in 2011, if Not as Blindingly Apple's sales w...
2011-01-03	Shares got a lift from the "January effect," when fund mana...	Wall Street Starts Year With a Surge	Wall Street Starts Year With a Surge Shares got a lift from th...
2011-01-03	Clean energy is a big moneymaker for Siemens, and the co...	Siemens Invests in Expanding Wind Power	Siemens Invests in Expanding Wind Power Clean energy is a...
2011-01-03	As part of its deal with Facebook, Goldman is expected to ra...	Goldman Offering Clients a Chance to Invest in Facebook	Goldman Offering Clients a Chance to Invest in Facebook As...
2011-01-03	In a second report, construction spending rose more than e...	Manufacturing Grew in December, but Housing Sector Strug...	Manufacturing Grew in December, but Housing Sector Strug...
2011-01-03	Fiat Industrial and Fiat's auto business met analysts' expecta...	Chief Says Fiat May Try to Add to Stake in Chrysler Before a ...	Chief Says Fiat May Try to Add to Stake in Chrysler Before a ...
2011-01-03	Mark Zuckerberg's dad | Detroit in ruins | Cuomo's photo str...	Morning Take-Out	Morning Take-Out Mark Zuckerberg's dad | Detroit in ruins |...
2011-01-03	Stamps that always remain valid, perks from vacation rentals...	Monday Reading: More Forever Stamps	Monday Reading: More Forever Stamps Stamps that always ...
2011-01-03	The following tax-exempt fixed-income issues are schedule...	Treasury Auctions for This Week	Treasury Auctions for This Week The following tax-exempt fi...
2011-01-03	Also in the news: the new schools chancellor's first day on t...	Morning Buzz | Cuomo Forecast Calls for Pay Freeze	Morning Buzz | Cuomo Forecast Calls for Pay Freeze Also in ...

Table A.1 Sample statistics of sentiment indices

Dataset	Loug		bing	
	Negative %	Positive %	Negative %	Positive %
NYT1	4.25	1.04	5.49	4.21
NYT2	4.48	1.21	5.81	4.30
NYT3	5.41	1.31	7.18	4.97
RWNC	6.40	0.50	8.71	2.61
Dow T	1.18	0.72	1.96	2.35
Google T	1.33	0.63	2.20	2.36
Apple T	1.51	0.78	2.80	2.69

Table A.2 Sample statistics for *Dow*

Date	Dow	
	Mean	St.dev
01.01.11-31.12.19	0.043	0.870
09.06.08-01.07.16	0.026	1.268
06.05.14-01.02.15	0.027	0.727

Figure A.2

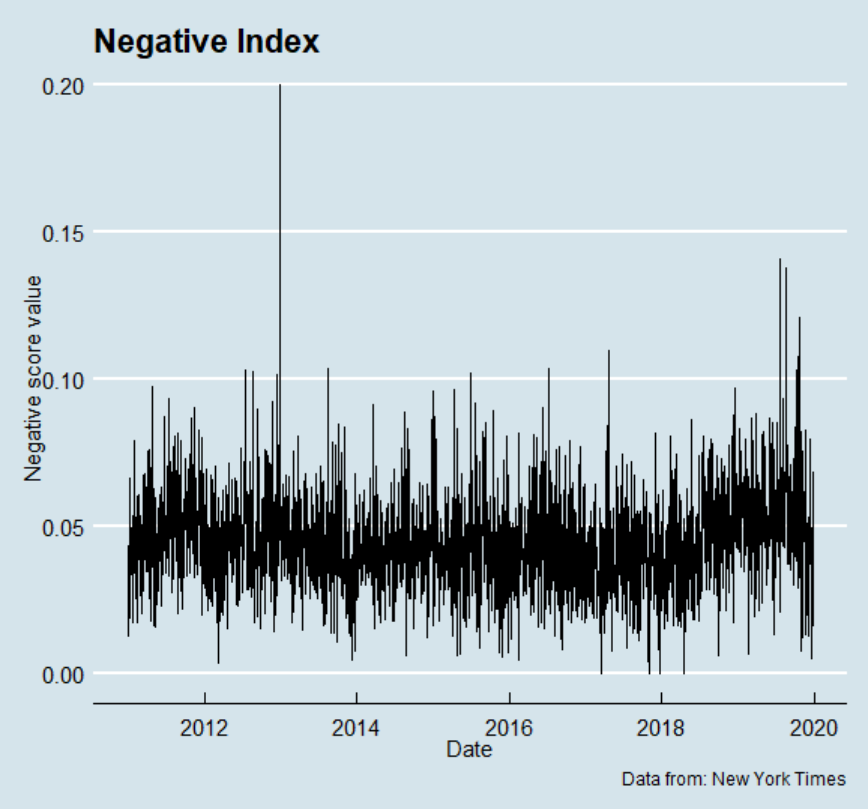


Figure A.3

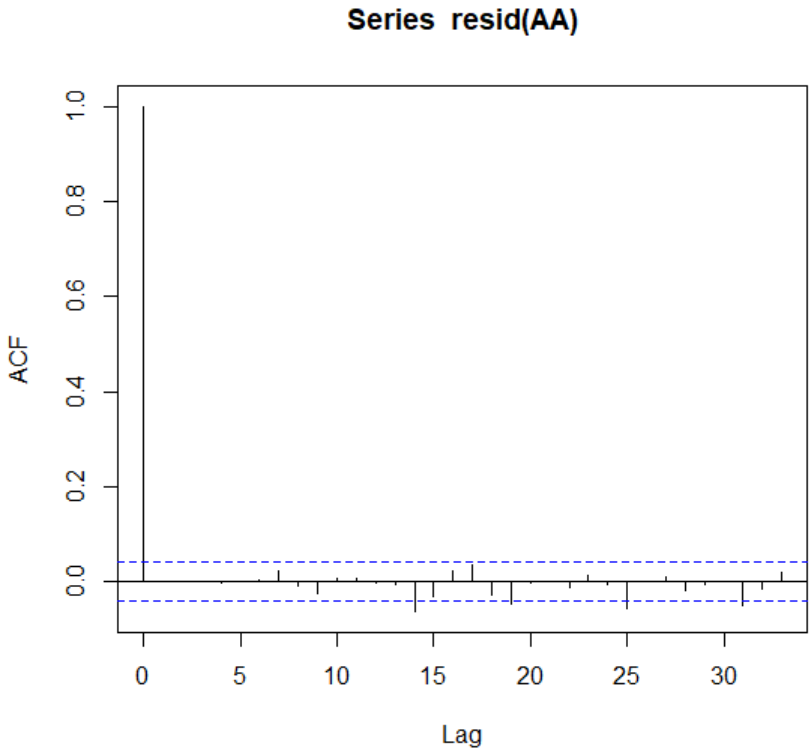


Figure A.4

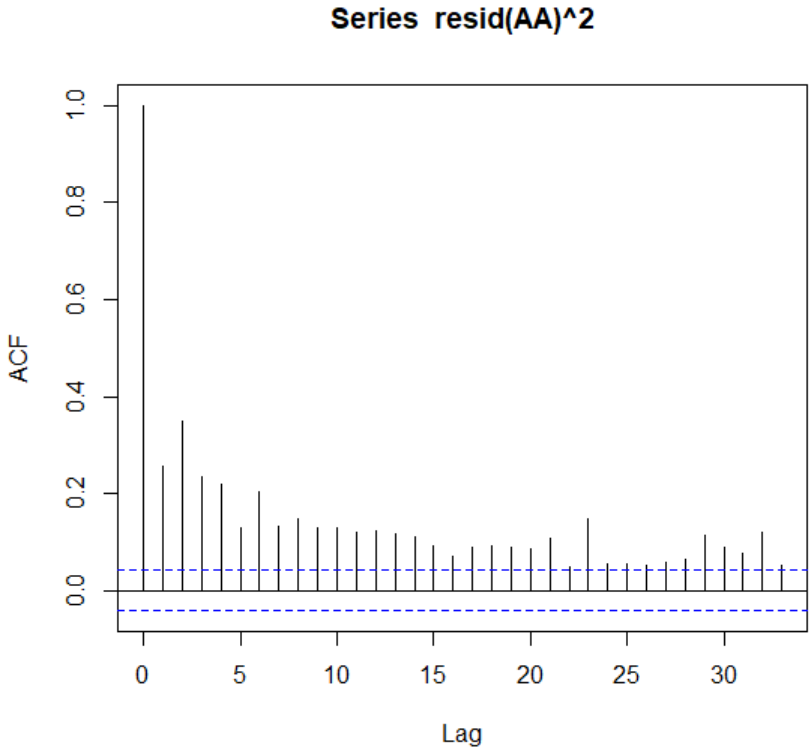


Table A.3 Loadings from PCA and PLS for the NYT2 dataset

	Dimensionality reduction loadings			
index/component	PC1	PC2	PLS1	PLS2
score_n	0.562	-0.061	-1.185	0.374
score_p	0.422	0.058	-0.158	0.188
score_l	0.463	-0.127	-0.224	0.962
score_s	0.069	0.988		
score_c	0.328	0.024		0.102
score_u	0.422	-0.018		

Table A.4 NYT1

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	-0.014	0.31	-0.012	0.39	0.020	0.64	0.016	0.35	-0.005	0.79
S_{t-2}	0.017	0.24	0.008	0.62	-0.085	***0.02***	-0.012	0.51	0.016	0.37
S_{t-3}	-0.013	0.36	-0.009	0.56	0.032	0.38	-0.002	0.90	0.002	0.90
S_{t-4}	0.006	0.64	0.003	0.82	-0.030	0.41	-0.010	0.55	0.000	0.99
S_{t-5}	-0.002	0.88	0.004	0.75	0.043	0.20	0.009	0.56	-0.012	0.43

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Table A.5 NYT3

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	-0.001	0.96	0.000	0.97	-0.005	0.86	-0.007	0.64	-0.020	0.18
S_{t-2}	-0.005	0.62	-0.019	**0.07**	-0.077	***0.01***	-0.002	0.90	0.020	0.13
S_{t-3}	0.014	0.11	0.016	0.11	-0.016	0.54	0.014	0.33	0.010	0.50
S_{t-4}	0.011	0.26	0.006	0.56	-0.042	**0.09**	0.009	0.55	0.017	0.28
S_{t-5}	0.010	0.36	0.016	0.17	0.013	0.62	0.014	0.38	-0.007	0.64

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Table A.6 Twitter Dow

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	0.048	0.63	0.063	0.54	0.033	0.81	-0.022	0.51	0.133	0.20
S_{t-2}	0.030	0.71	0.119	0.23	0.176	0.27	-0.035	0.37	0.032	0.72
S_{t-3}	-0.016	0.86	0.005	0.96	0.058	0.66	0.007	0.77	-0.191	0.12
S_{t-4}	-0.020	0.82	-0.153	0.16	-0.267	0.13	0.057	0.20	-0.052	0.55
S_{t-5}	-0.012	0.90	-0.044	0.62	-0.055	0.63	0.016	0.59	0.068	0.33

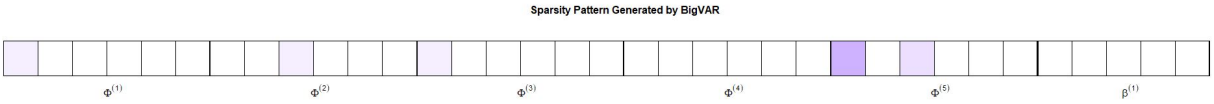
The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Table A.7 Apple

	Score		N		P		PC		PLS	
	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val	ϕ	p-val
S_{t-1}	-0.215	0.50	-0.156	0.48	-0.119	0.70	0.066	0.47	-0.041	0.92
S_{t-2}	0.216	0.29	0.125	0.50	0.048	0.88	-0.093	0.30	0.304	0.44
S_{t-3}	-0.160	0.51	0.058	0.77	0.462	**0.10**	-0.080	0.25	-0.364	0.25
S_{t-4}	0.390	0.14	0.260	0.34	-0.015	0.97	-0.049	0.56	0.535	**0.08**
S_{t-5}	-0.397	***0.02***	-0.390	***0.04***	-0.462	0.21	0.194	0.04	-0.265	0.57

The p-value are reported using Newey and West standard errors, adjusting for heteroskedastic and autocorrelation for five lags. Bold indicate significance on a 10% level. Bold and italic indicate significance on a 10% level.

Figure A.5 Illustration of penalized parameters for regression with NYT3 text data



The intencity of the color discribes the penalty, white parameters have been shrunken to 0.

Figure A.6 Penalized parameters values for regression with NYT3 text data

```
> VARf@betaPred
          [,1]         [,2] [,3]        [,4] [,5] [,6] [,7] [,8] [,9]        [,10] [,11] [,12] [,13]        [,14] [,15]        [,16] [,17] [,18] [,19]
[1,] -0.00123632 -0.006524715    0 0.00000000    0    0    0    0    0 -0.001302791     0     0     0 -0.001940299     0 0.000000000     0     0     0
[2,]  1.34073891  0.000000000    0 0.00000000    0    0    0    0    0 0.000000000     0     0     0 0.000000000     0 0.000000000     0     0     0
[3,]  4.25749892 -0.005552671    0 0.09485956    0    0    0    0    0 0.121459784     0     0     0 0.000000000     0 0.002775796     0     0     0
[4,]  0.83292046  0.000000000    0 0.00000000    0    0    0    0    0 0.000000000     0     0     0 0.000000000     0 0.000000000     0     0     0
[5,]  0.65230156  0.000000000    0 0.00000000    0    0    0    0    0 0.000000000     0     0     0 0.000000000     0 0.000000000     0     0     0
[6,]  0.32695357  0.000000000    0 0.00000000    0    0    0    0    0 0.000000000     0     0     0 0.000000000     0 0.000000000     0     0     0
     [,20] [,21] [,22] [,23] [,24] [,25]        [,26] [,27]        [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36]
[1,]     0     0     0     0     0     0 -0.04175356     0 0.008929144     0     0     0     0     0     0     0     0
[2,]     0     0     0     0     0     0  0.00000000     0 0.003294011     0     0     0     0     0     0     0     0
[3,]     0     0     0     0     0     0  0.00000000     0 0.000000000     0     0     0     0     0     0     0     0
[4,]     0     0     0     0     0     0  0.00000000     0 0.000000000     0     0     0     0     0     0     0     0
[5,]     0     0     0     0     0     0  0.00000000     0 0.000000000     0     0     0     0     0     0     0     0
[6,]     0     0     0     0     0     0  0.00000000     0 0.000000000     0     0     0     0     0     0     0     0
```

The figure shows the penalized parameters

R-libraries

- *library(tidyverse)*
- *library(dplyr)*
- *library(tidytext)*
- *library(jsonlite)*
- *library(readr)*
- *library(httr)*
- *library(magrittr)*
- *library(fansi)*
- *library(zoo)*
- *library(textdata)*
- *library(lmtest)*
- *library(aTSA)*
- *library(ggthemes)*
- *library(rugarch)*
- *library(vars)*
- *library(mltools)*
- *library(fastDummies)*
- *library(forecast)*
- *library(FactoMineR)*
- *library(factoextra)*
- *library(lubridate)*
- *library(aTSA)*
- *library(sandwich)*
- *library(pls)*
- *library(BigVAR)*
- *library(mice)*
- *library(rtweet)*