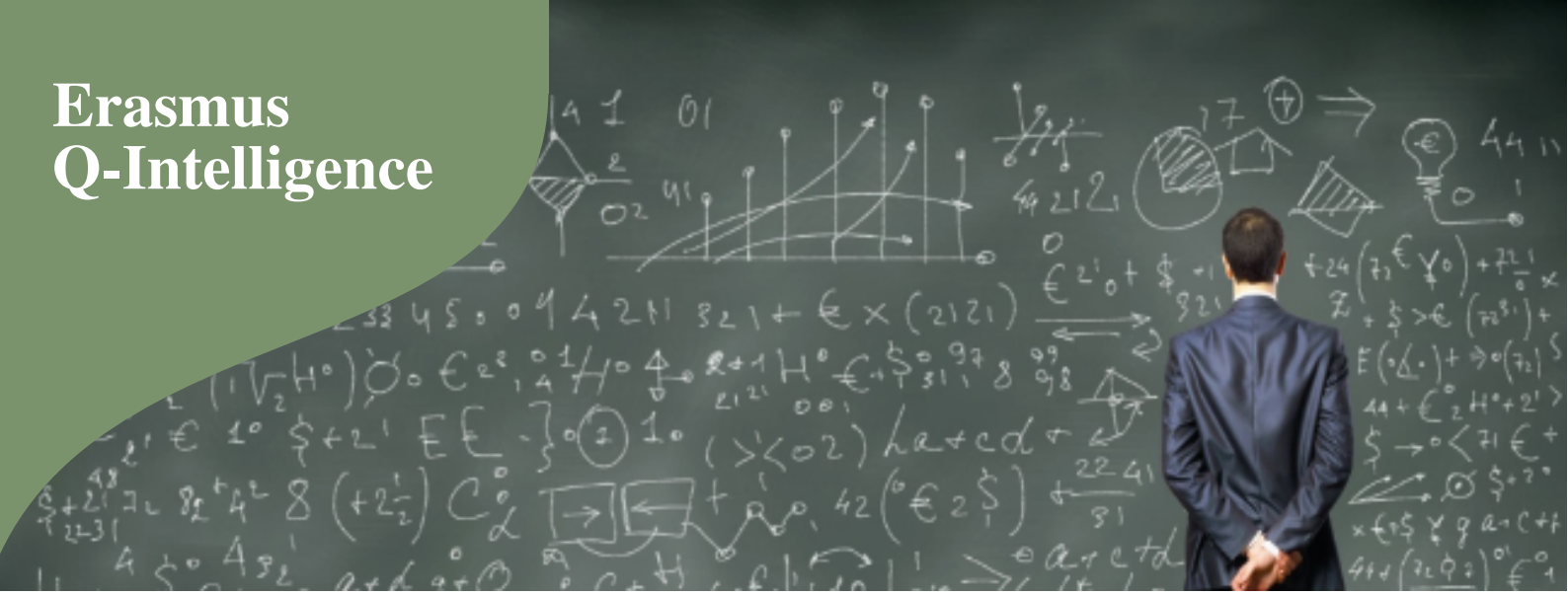
Erasmus
Q-Intelligence



Forecasting Dutch GDP

Incorporating Sentiment from Dutch News Articles

Master Thesis

Martin Skogholt

Supervisor: Kristiaan Glorie
Michel van de Velden
Co-reader: Kim Schouten

ERASMUS UNIVERSITEIT ROTTERDAM

Abstract

To facilitate stable economic development and prevent bankruptcy both business- and governmental institutions need reliable estimates of current GDP. In the Netherlands, however, official GDP estimates have a significant publication lag of 45 days after the end of a quarter. The EICIE model developed by de Groot and Franses (2006) made the first step by using staffing data from a temp agency. This data is available a week after the end of a quarter. To improve the predictive accuracy of EICIE, a novel approach to utilize news articles by applying sentiment analysis techniques is developed in this thesis. The approach reduces the publication lag while improving the accuracy of the estimates. Moreover, some dimensionality reduction and variable selection methods have been shown to considerably boost predictive performance, when faced with the sparse data sets generated by the approach adopted in this thesis. The prediction accuracy is significantly improved upon with a 65.4% improvement compared to the EICIE model if both models are re-estimated after each forecast and an 81.6% improvement in prediction accuracy if the models are only estimated once with the training data.

Keywords: Sentiment Analysis, GDP Forecasting, News Articles.

Contents

1 | Introduction

A recession does not have an official definition, but a practical definition of a recession can be given as a decline of a country's real GDP for two or more consecutive quarters according to Claessens and Ayhan Kose (2009). A recession is characterized by plummeting sales, crashing housing markets, soaring unemployment, and a myriad of other negative effects. Despite these adverse effects, recessions are an inherent part of the economic cycle. There have been 122 recessions in 21 advanced economies from 1960-2007 according to Claessens and Ayhan Kose. The latest recession in the Netherlands is the one that started in 2008. Housing prices plummeted, unemployment surged, and businesses went bankrupt left and right.

The severity of a recession can be reduced with interventions by policymakers and government institutions. Blinder and Zandi (2010) showcase the different measures that were taken by the U.S. to minimize the impact of the recession of 2008. Businesses can also intervene and take steps to avoid bankruptcy and to retain stable sales and profits. According to Kitching et al. (2009), there is a range of different strategies that can be employed by businesses to not only survive a recession but to prosper during an economic downturn. Consequently, policy makers and businesses need to know the present state of the economy to enable stable economic development.

According to Callen (2008), Gross Domestic Product(GDP) is the most widely used gauge of economic health. GDP is cited in reports by central banks, business communities, and various governmental bodies. GDP is the value of all the goods and services that are produced in a country or region in a certain time period (i.e. in a quarter or in a year).

In many countries, official GDP estimates are only published on a quarterly basis and often with a significant delay. Statistics Netherlands(SN)[1] publishes Dutch GDP estimates. The first GDP estimate, which is published 45 days after the end of a quarter, is a so-called flash estimate. The UK is the fastest of major economies to publish an official GDP estimate, but still, the estimates are published 25 days after the end of a quarter[2].

Not only is there a substantial delay in the publication of GDP estimates, but the reliability of these GDP estimates is also an issue. At the point in time of the first estimate, only around 40% of the data used to produce GDP estimates is available. These estimates are updated as more data becomes available and this is the reason for the frequent revisions of GDP publications. In the Netherlands, SN updates the initial flash estimate 45 days later, which is 90 days after the end of the quarter in question, to a so-called Regular Quarterly Forecast. These estimates are adjusted in July in the following year, a total of 1 to 5 quarters later than the quarter for which GDP has been estimated. These are called the adjusted Regular Quarterly Forecasts. One year after this, a 'preliminary definite' GDP estimate is given, and another

[1] https://www.cbs.nl

[2] http://www.bbc.com/news/business-13200758

year later the 'final definite' value of the Dutch GDP is published. Still, even this final estimate can be revised. Undoubtedly, Knowing the present state of the economy with a substantial lag does not make it easy for companies and policy makers to react promptly and adjust policies to intervene.

Due to the untimely publication of GDP, there have been many studies on how to forecast GDP of the current quarter, the most recently ended quarter, or GDP of future quarters. Most of the research that has been done on forecasting GDP focuses on using economic indicators or on which statistical model to use. For example, den Reijer (2005) used a data set of 370 economic variables to forecast Dutch GDP. Using this data set led to an increase in prediction accuracy of up to 30% compared to a benchmark model without any economic variables.

A previously untapped source of information is unstructured textual data available online. There is a large amount of information readily available that could reflect the current state of the economy and consumer confidence. Exploiting this information could potentially improve the prediction accuracy of existing models that incorporate a multitude of economic indicators. Furthermore, using the data available on the web could potentially effectuate GDP estimates virtually instantaneously after the end of a quarter, if the predictive accuracy is adequate with only the data available on the web. This leads to the question whether the predictive performance of existing models can be improved upon by using online data and whether the extracted data could successfully be used without any additional economic indicators to eliminate any publication lag.

In order to investigate the benefit of using online information, a novel framework to extract and quantify information from unstructured online sources is developed in this thesis and evaluated on the predictive performance achieved. In order to quantify textual information, sentiment analysis techniques are applied. These techniques have two tasks; one is to determine the sentiment in a sentence and the second is to determine which word in a sentence the sentiment is about. A sentiment is an opinion about something. In this thesis, the sentiment is denoted by either +1 (positive) or -1 (negative). After determining the sentiment with regard to selected words related to economic development, the sentiment scores are aggregated for each of the selected words to denote the sentiment about these words in a certain time period, for example, the first quarter of 2005. This means that each of the selected words with regard to economic development has a corresponding time series variable that denotes the sentiment about this word throughout the sample period.

There are many selected words with regard to economic development, and as a result, the number of variables is large (i.e. $p \gg n$). Most statistical models cannot cope with such a large number of variables and hence dimensionality reduction and variable selection techniques are implemented to reduce the number of variables. The reduced variables are incorporated in a baseline model called "EICIE", which was developed by de Groot and Franses (2006). Furthermore, machine learning models are utilized to forecast Dutch GDP with both the reduced set of variables, as well as the original set of variables. These models are designed to handle a large number of variables in a sensible way.

The remainder of this thesis is organized as follows: chapter 2 describes literature on GDP estimation models and current adaptations of online sources for economic forecasting. Additionally, literature on sentiment analysis techniques is discussed. In chapter 3, the methodology of the developed framework and its components is discussed in detail. In chapter 4, the results of the developed framework and each individual component are presented. Finally, a conclusion is given and directions for future research are addressed in chapter 5.

2 | Literature

In this chapter, an overview of the current literature is presented. In section 2.1, previous work on GDP estimation models incorporating economic variables and indicators is discussed. In section 2.2, an overview of the current adaptations of online sources towards forecasting economic variables is given. In section 2.3, the methods of sentiment analysis that serve as the basis of our sentiment determination method are described. Finally, some papers are presented that form the basis of how relevant words related to economic growth are determined in section 2.4.

2.1 | Previous Work on GDP Estimation

De Groot and Franses (2005) propose to use temporary staffing data[1] to forecast the Dutch GDP. De Groot and Franses hypothesize that temporary staffing activity is closely linked to economic growth and temporary staffing data is available quickly after the end of a quarter. They construct a model for the yearly growth rates of Dutch GDP, as well as for the quarterly growth rates of GDP. Both models contain autoregressive components and temporary staffing data, both static and dynamically. De Groot and Franses further show that there is a co-integration relationship between the two time series in the

[1]Provided by Randstad

long run. In de Groot and Franses (2006) they extend the model previously mentioned by also investigating the presence of cycles in the data by including harmonic regressors. De Groot and Franses use a bivariate vector error correction model with harmonic regressors to model the Dutch GDP. However, they show that the harmonic regressors are all insignificant for cycle lengths of 1 to 15 years. They use the resulting error correction model, which is called "EICIE", to produce forecasts up to Q4 2015.

A disadvantage of both aforementioned models is that they cannot incorporate monthly economic indicators with various publication lags to arrive at earlier forecasts. For example, after the first month, a preliminary forecast could be made for the current quarter, which is then updated as new information is published. There has been some research into using more sophisticated models that incorporate monthly economic indicators to produce quarterly GDP Forecasts. Baffigi et al. (2002) proposed to use bridge models to model GDP of the Euro area and three selected European countries and show that these models outperform the univariate models, such as the ARIMA, VAR, and structural models. These bridge models essentially 'bridge' monthly data to quarterly GDP values.

A disadvantage of a bridge model is that it can only contain a limited number of regressors. Since there are many economic indicators, Giannone et al. (2008); Marcellino and Schumacher (2010); Schumacher and Breitung (2008) have proposed an extension to the bridge models to be able to handle a multitude of regressors. These so-called factor models implement Principal Com-

ponent Analysis(PCA) to summarize a large number of economic variables in a few components and subsequently use these components to forecast GDP. These factor models have been shown by both Angelini et al. (2011) and Rünstler et al. (2009) to outperform the bridge models. Den Reijer (2005) implemented a dynamic factor model to forecast the Dutch GDP up to 8 quarters ahead. Den Reijer used a subset out of 370 macroeconomic variables in the factor matrix and show that the dynamic factor model outperforms a standard autoregressive benchmark model, however, only with the optimal subset of variables. The gain in prediction accuracy ranges from 10% to 30% with respect to a benchmark model, dependent on the selected subset.

The EICIE model serves as the baseline model in this thesis. The variables, generated from news articles, are added to the EICIE model in an effort to improve the predictive accuracy. The idea of applying PCA to reduce the number of variables is used as a dimensionality reduction technique. The variables that are generated from news articles could also have been added to a bridge model or one of the aforementioned factor models to effectuate monthly forecasts of the current GDP, however, this is considered to be outside the scope of this thesis.

2.2 | Incorporating Sentiment to Forecast GDP

A large number of textual sources are available on the web, such as news articles, blogs, social media, etc. In this thesis, the use of news articles is exhibited. The developed framework can, however, easily be applied to other sources, such as Twitter or blogs.

There has been some research on using news articles to increase the performance of GDP forecasts. One of these is a study conducted by Kourentzesa and Petropoulosa (2014), which uses the sentiment of newspaper articles about the current economy to improve the accuracy of GDP nowcasts. In order to use information from news articles to forecast GDP, Kourentzesa and Petropoulosa create new variables by using the distances between words in news articles to create word pairs that capture some sentiment about the economy. They do not determine the actual opinion or sentiment, but rather use the frequency of the word pairs in news articles as numerical variables to forecast GDP. Each word pair generates an individual variable with the frequency that the word pair occurs in the news in a certain time period. This is also the limitation of this study. Kourentzesa and Petropoulosa do not determine the sentiment of the news articles, but implicitly infer this by using the correlation between the frequency of the word pairs and GDP growth. Additionally, the words initially used to generate the word pairs are arbitrarily chosen and may not be the best words to use for this purpose.

Bozic and Seese (2011) use the Reuters NewsScope Sentiment Engine to retrieve sentiment scores of news articles and simply show that these sentiment scores have a significant influence when incorporated in a model of GDP growth. They use the sentiment scores aggregated over the entire year for a large number of countries. It is not clear how the Reuters NewsScope Sentiment Engine assigns the sentiment scores, but it is an indication that sentiment from news articles can be useful in predicting GDP. A disadvantage of this approach is that the sentiment scores are for the news article as a whole. A news article might contain conflicting sentiments about the same aspect or multiple sentiments concerning multiple aspects. An aspect is the target of an opinion. Remember, an opinion is always about a certain something and this something is called an aspect in this thesis.

Tuckett et al. (2015) also studied how to incorporate news articles sentiment to increase the accuracy of GDP estimates. They categorize the sentiment in two categories: "excitement about gain" and "anxiety about loss". Tuckett et al. created two dictionaries of words, one for each category, with pre-selected English words that convey or evoke the aforementioned emotions. They subsequently specify a relative sentiment shift (RSS) as the relative strength of the two emotions present in a text database from the Thompson Reuters news archive from 1996 till 2014. Tuckett et al. counted the number of words in the text that belong to the two categories and define the RSS as the difference between the two groups, divided by the size of the text corpus. In other words,

equation 2.1 is used to calculate the RSS at time t.

$$RSS_t = \frac{E_t - A_t}{N_t} \tag{2.1}$$

Where N_t is the size of the text corpus at t, E_t is the number of words belonging to the "excitement about gain" group at t, and A_t is the number of words belonging to the "anxiety about loss" group at t. Using equation 2.1 a time series variable is generated that can be used to forecast GDP. The downside is that these emotions can be expressed about anything and that the aspect of the sentiment is not used. Knowing not only the sentiment itself but the aspect of the sentiment could lead to an improvement in accuracy.

To the best of our knowledge, there is no other research on incorporating news sentiment to forecast GDP. In the above papers, either the general sentiment is determined, but not the aspect of the sentiment, or the aspects are indirectly determined, but not the sentiment itself. The developed framework in this thesis improves upon this by doing both.

2.3 | Extracting Sentiment

Determining the sentiment in news articles to forecast GDP is, as of yet, not a widely researched area, but sentiment analysis of customer reviews has been researched extensively. Sentiment analysis techniques successfully applied to customer reviews are used as a basis for the sentiment analysis determination method developed in this thesis.

Qiu et al. (2011) propose a so-called "semi-supervised sentiment aspect extraction method". It is an algorithm to detect the aspects of a certain opinion in the text and to expand a "sentiment word" dictionary. A "sentiment word" is a word that conveys a certain sentiment, such as bad, great, horrific, terrific, etc.. Qiu et al. extract the aspects by employing a set of rules that use the syntactic relationships in a sentence. A syntactic relationship is a relationship between two words in a sentence that gives a certain structure to the sentence. For example, a verb in a sentence often has a subject. This means that the verb and a noun in the sentence have the syntactic relationship called subject. The algorithm, developed by Qiu et al., only needs a set of known sentiment words to work and no further data is required. Based on the rules, a so-called double propagation effect occurs, where sentiment words are used to extract aspects and the extracted aspects are used to find new sentiment words that were not already present in the initial dictionary. The procedure is executed iteratively until no new sentiment words or aspects can be found. In this way, not only aspects are extracted, but also new sentiment words are found. The polarity (positive or negative) of these new sentiment words can then be classified based on the polarity most frequently associated with the aspect with which the new sentiment word was found. The assumption is that in reviews (which is the domain of their research) the polarity of a certain aspect will stay the same throughout the review since it is written by the same author.

Qiu et al. employ a total of 8 handcrafted rules, using the syntactic relationships, to extract the aspects and expand the sentiment word dictionary.

To illustrate the use of syntactic relationships to extract aspects, consider the following sentence:

The Netherlands has a great economy.

Here, "great" is a known sentiment word. In order to apply the rules using the syntactic relationships, the sentence first needs to be parsed. Parsing is the analysis of a sentence to determine the syntactic relationships between the words in a sentence and the Part-of-Speech(PoS) of each word. Parsing results in a so-called parse tree that contains these syntactic relationships between all of the words in the sentence along with Part-of-Speech tags. A Part-of-Speech(PoS) is a category of words, which have similar grammatical properties. For example, "noun", "verb", "adjective". The parsed sentence is displayed in figure 2.1. Here, the Part-of-Speech(PoS) tags are depicted below the words themselves and the syntactic relationships are marked as the blue text.

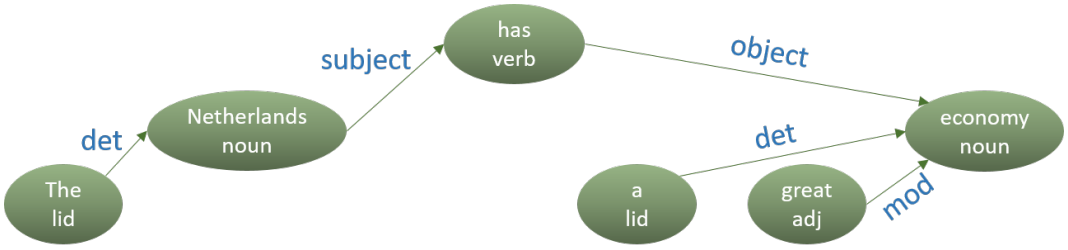


Figure 2.1: An example of a parsed sentence. The PoS tags are in the globes under the word and the syntactic relationships are in blue

As can be seen, "great" is an adjective and it is the known sentiment word. It has a "mod" relationship with the noun, "economy". Using the first rule

of Qiu et al., "economy" is classified as an aspect. Additionally, "economy" is related to "Netherlands" through "subject" and "object". Hence, according to the second rule of Qiu et al., "Netherlands" is also classified as an aspect. There are six more rules employed by Qiu et al. to extract aspects and find new sentiment words to expand the dictionary.

A disadvantage for this thesis is that these syntactic rules are developed for the English language and might not apply to Dutch. Additionally, the research was performed using reviews, which sole purpose is to convey a sentiment about a certain product. News articles, however, are more informative than subjective and the rules developed by Qiu et al. might not apply to news articles.

A popular method to determine the sentiment in a sentence is to use a voting scheme in conjunction with a known sentiment lexicon. If there are more positive sentiment-bearing words than negative ones, the sentiment is defined as positive, or +1, and vice versa. In the case of a tie, the most frequent sentiment in the data is assigned as the sentiment. Hu and Liu (2004); Moghaddam and Ester (2010); Zhu et al. (2009) employed this method successfully and Choi and Cardie (2008) show that this simple scheme achieves an accuracy of 87.7% in classifying the sentiment in the MPQA data set[2].

Zhu et al. (2009) notes that sentences can carry multiple opinions (i.e. a combination of a sentiment and an aspect) in a single sentence and propose a method to segment sentences with multiple aspects to so-called sentiment-bearing segments. This method is derived from different scores on how important a certain

[2]http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

extracted aspect is. The method is unsuitable for our purposes since it needs labeled training data, but it does show a 5% increase in correctly classified opinions when taking multiple opinions in a single sentence into account.

In this thesis, a novel sentiment determination technique is implemented, based on the voting scheme of Choi and Cardie and the use of syntactic relationships, but given that the aspect is already known. The usage of syntactic relationships makes it possible to know which sentiment word is about which aspect in sentences with multiple opinions.

2.4 | Determining Aspects

Vosen and Schmidt (2011, 2012) introduced a new indicator for forecasting private consumption of the US and Germany, respectively. Since GDP is the sum of private consumption, government spending, investments, and net exports, this seems useful also for GDP. Vosen and Schmidt used internet search query data obtained from Google to forecast private consumption. Google delivers internet search query data in different categories and subcategories. Vosen and Schmidt manually selected certain consumption related categories and used principal component analysis to reduce the number of variables. The extracted components were then used in an autoregressive model to forecast the private consumption. These components improved the out-of-sample forecasting performance of the model compared to models with different consumer confidence indicators.

Choi and Varian (2012) used Google's internet search query data to predict economic indicators on automotive sales, unemployment claims, travel destination planning, and consumer confidence. Choi and Varian go on to show that using Google's search query data improves an autoregressive baseline model up to 20% when also incorporating the internet search query data. This result is another indication that internet search query data could be useful for forecasting present GDP.

As previously explained, a procedure to determine relevant aspects is still needed. Using Google search query data has been proven to be useful in forecasting GDP and hence the top search queries are going to be used as the relevant aspects. The aggregated sentiment with regard to these relevant aspects is used as the variables, instead of using the search volume as numerical variables.

3 | Method

In this chapter, the developed framework to extract and quantify textual information from news articles is presented in detail. The sentiment with regard to a list of aspects has to be determined, which is elaborated upon in section 3.1. Our framework generates a large number of variables. To handle this large number of variables, some variable selection and reduction methods are used, as discussed in section 3.2. After a selection or reduction of the variables is made, the resulting variables are used in existing models to forecast GDP, which are described in section 3.3. The models that are used to forecast GDP have a number of parameters that need to be optimized. In section 3.4, the parameter optimization procedure is presented. Finally, in order to optimize these parameters, the used measures of performance are explained in section 3.5.

3.1 | Sentiment Determination Procedure

A first step is to collect the news articles from the first quarter of 2000 until the end of 2015. This a trade-off between processing time and accuracy. A longer time sample should increase the accuracy and lead to more stable and robust coefficient estimates. Dutch GDP is only published quarterly by SN, while the number of news articles is large at around 50 articles each day.

This means that the duration of processing all these news articles increases rapidly for each additional quarter. The collection is done by building a custom web scraper to collect all the news articles from the online archive of the "Volkskrant"[1]. Not all Dutch newspapers have an online archive that is suitable for our purposes, but the Volkskrant is the largest newspaper with an archive suitable for scraping. The articles are filtered by their category assigned by the Volkskrant. Only so-called "opinie" articles are used in this thesis. This category is the same as editorials or opinion pieces. The reason for this is that this type of news articles will most likely contain the most opinions to extract. From this online archive, around 2 million articles are collected and processed. In subsection 3.1.1, the determination of the relevant aspects is explained. In subsection 3.1.2, the actual sentiment classification procedure is presented and finally, in subsection 3.1.3, the sentiment aggregation procedure is elaborated upon.

3.1.1 | Extracting Relevant Aspects

As was previously mentioned, a sentiment is always about a certain something, called an aspect in this thesis. This means that aspects have to be defined to extract the sentiment about these aspects from news articles. In order to extract relevant aspects that can be used in the news sentiment determination procedure, Google Trends data and phrases from the Finler Encyclopedia[2] are

[1] http://www.volkskrant.nl/

[2] https://www.finler.nl/

used. The two sources, combined, yield 3993 aspects. The sentiment about these aspects represents both the consumer sentiment, as well as the sentiment about the financial and economic activity.

Google has categorized all search queries into a number of categories and subcategories. For each of these, a list of top search queries can be found online. All of these top search queries, for each category, are collected and serve as our aspect list. Intuitively, the search queries that have a large search volume are deemed important by consumers and the sentiment about these queries might represent consumer confidence. Since a large part of the GDP of a country is contributed to consumer spending, these variables could potentially contain relevant information for forecasting GDP.

In addition to using Google's search data, a list of over 1500 aspects from the Finler Encyclopedia[3] of economic and financial terms is extracted. The sentiment with regard to these financial and economic terms give an indication as to how the economy is developing. For example, if there is negative sentiment about the stock market, this might be an indication that the stock market is going down and that economic activity is decreasing.

3.1.2 | Sentiment Classification

The actual sentiment classification procedure is based on the voting scheme of Choi and Cardie. To use this scheme, a list of sentiment words is needed.

[3] https://www.finler.nl/

The number of occurrences of known sentiment words is counted for a single sentence and the most frequent polarity is assigned as the dominant sentiment. For example, if two sentiment words are known to be positive and a single sentiment word known to be negative, then the sentiment is classified as positive. In the case of a tie, the most frequent encountered polarity in the rest of the text is assigned as the sentiment. This means that a Dutch sentiment word list is needed. For this purpose, the subjectivity lexicon developed by De Smedt and Daelemans (2012) is used.

The collected news articles are processed to filter out sentences that do not contain a combination of a known aspect and a known sentiment word from the sentiment dictionary and the aspect list, respectively. For each of the remaining sentences, the sentiment needs to be determined. The procedure to determine the sentiment is inspired by the voting scheme of Choi and Cardie and the method of Qiu et al.. The rules, employed by Qiu et al., are for extracting aspects and sentiment words; however, in this case, the aspects are already known. The idea of parsing sentences and using the syntactic structure of a sentence is implemented in the new sentiment determination procedure developed in this thesis. In order to determine the syntactic structure, the Alpino Parser, developed by Bouma et al. (2001) is used. The idea is that each known aspect has a syntactic path to the sentiment word in the corresponding sentence. The syntactic path consists of a number of syntactic relationships that are traversed from the aspect word to the sentiment word. The Alpino Parser produces 11 syntactic relationships. A list of the syntactic relationships

with an example as to how the syntactic relationships occur in a sentence is displayed in table 3.1. Practical definitions of each syntactic relationship are displayed in table 3.2.

Table 3.1: A list of the different syntactic relationships that were used and an example of these syntactic relationships

Syntactic Relationship	**Example**
Subject	The telephone is great
Mod	The telephone has a great screen
Object1	The telephone has a screen
Object2	Shall I pour a drink for you?
Predc	The telephone is beautiful
Cnj	The telephone is big and beautiful
PC	He is accused of fraud
VC	He is accused of fraud
DP	Stranded tourists home again
Nucleus	If you win, then you have won
Sat	If you win, then you have won

The syntactic structure of a sentence can subsequently be used to determine the sentiment in a sentence. For example, consider again the sentence with a parse tree displayed in figure 2.1:

> The Netherlands has a great economy.

Additionally, let "Netherlands" and "great" be a known aspect and a known sentiment word, respectively. In this case, the syntactic path between the two is "sub - obj1 - mod". From now on, the composition of an aspect, a sentiment word, and a syntactic path, such as "Netherlands, great, sub - obj1 - mod" will be referred to as an opinion. Opinions, like the above, are extracted from every relevant sentence. In total, this results in approximately 2 million opinions. In

Table 3.2: Syntactic Relationships and their description.

Syn.	Description
Subj	The subject is the word or phrase which controls the verb in the clause, that is to say with which the verb agrees.
Mod	A modifier is said to modify or change the meaning of another element in the structure, on which it is dependent.
Obj1	Object1 is the direct object in a sentence, which is the entity that is acted upon by the subject.
Obj2	Object2 is the indirect object in a sentence, which is an entity that is indirectly affected by the action.
Predc	The predicate is an entity that completes an idea about the subject, such as what it does or what it is like.
Cnj	A conjunct is the relation between two elements connected by a coordinating conjunction, such as “{”and}, “{”or}, etc.
PC	This is used when the complement of a preposition is a clause or prepositional phrase (or occasionally, an adverbial phrase).
VC	Verb phrase is a syntactic unit composed of at least one verb and its dependentsâĂŤobjects, complements and other modifiers
DP	This is used for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way)
Nucl	This is an independent clause that can express a complete thought (and can be a standalone sentence).
Sat	This is a dependent clause that is usually a supporting part of a sentence, and it cannot stand by itself as a meaningful proposition (idea).

the above example, the opinion with “economy, great, mod” would have also been extracted.

The sentiment about an aspect is the same as the polarity of the sentiment word to which the aspect is linked by a syntactic path. This is similar to the voting scheme, as described in Hu and Liu; Moghaddam and Ester; Zhu et al.; Choi and Cardie. The major difference is that aspects are linked to sentiment words based on the syntactic path. This means that aspects are linked to sentiment

words in accordance with the structure of the sentence, rather than treating the sentence as a single whole, such as in the voting scheme. Furthermore, it allows putting restrictions on the opinions that are considered to be correct. A restriction can be placed on the total length of the syntactic path, or on which syntactic relationships that are allowed. For example, the restriction can be that no opinions are allowed with a "mod" relationship. In this case, both the above opinions would be discarded from any further analysis.

In the case that there are no restrictions on the path, a subtle difference arises compared to the voting scheme. If there are, for example, three sentiment words in a sentence, the sentiment about an aspect would be classified as the most common polarity of these three sentiment words, when using the voting scheme. For example, if the are two positive and one negative sentiment word, the sentiment would be determined as positive. Consider the sentence:

> The iPhone has a great, beautiful, but dirty screen.

In this case, using the voting scheme would result in a single opinion consisting of "screen" with a positive sentiment. In our case, three opinions are generated: "screen, great, mod - cnj", "screen, beautiful, mod - cnj", "screen, dirty, mod - cnj". This gives a finer distinction in sentiment classification than simply positive or negative. Moreover, if there are restrictions on the syntactic paths, one or more opinions might be discarded, which would result in a different aggregated sentiment entirely. This allows a big degree of flexibility in our sentiment determination procedure compared to the voting scheme. Moreover, if there is a sentence with two aspects and two sentiment words and the sentiment words

are of contradictory polarity, one positive and the other negative, such as:

The school is great, but the kitchen is dirty.

The voting scheme would classify one of the opinions incorrectly. The reason is that the voting scheme classifies the sentiment for both these aspects as the most frequent polarity in the text since there is a tie between the two sentiment words. Either "school" and "kitchen" would be determined to both have a positive sentiment or both a negative sentiment. The more probable situation is that one aspect belongs to one sentiment word and the other aspect to the other sentiment word, such as "school" with a positive sentiment and "kitchen" with a negative sentiment, which is the case in this example. With our approach, this is allowed and will be properly classified dependent on the restrictions on the syntactic path. In fact, our procedure will classify four opinions, all the combinations of an aspect and a sentiment word that occur in the sentence. With the restrictions on the syntactic paths the developed framework can be able to discard the wrong combination of aspect and sentiment word and classify this as "school" with a positive sentiment and "kitchen" with a negative sentiment. Even more so, the developed framework can classify the sentence as "school" with a positive sentiment, "kitchen" with a negative sentiment, and "school" with a negative sentiment. This is the best situation, since the "kitchen" is the kitchen in the school and since it is dirty, there is also an implicit negative opinion about the school. At least, as far as schools with clean kitchens are preferred.

3.1.3 | Generating Sentiment Variables

As previously explained, opinions are generated that consist of an aspect, a sentiment word, and a syntactic path. It is possible to put restrictions on the syntactic paths. There are 11 syntactic relationships that are found with the Alpino Parser. Also, each path has a certain length. All these combinations of restrictions generate different sentiment variables to be used for predicting GDP. All of the extracted opinions need to somehow be converted to a score and aggregated on the relevant time periods. In our case, the sample starts at the beginning of 2000 and ends at the end of 2015. Moreover, it consists of quarterly observations of GDP. This means that the opinions have to be aggregated on a quarterly basis. All of the opinions consists of an aspect and hence the sentiment with regard to a certain aspect over a time period can be determined. To aggregate the opinions, a Relative Positivity Score (RPS) is defined in equation 3.1:

$$RPS_t = \frac{\#\ Positive\ Opinions_t - \#\ Negative\ Opinions_t}{\#\ Opinions_t} \tag{3.1}$$

This score represents the positivity about an aspect in quarter t. This score is calculated for all of the extracted aspects over the whole sample and results in a data matrix, where the column represents an aspect and the row represents a quarter. Since our sample ranges from 2000Q1 to 2015Q4, this results in a $[64 \times \#\ Aspects]$ matrix. Remember that there were 3993 possible aspects and hence the resulting data set has a large number of variables.

3.2 | Variable Selection and Reduction

One option to reduce the dimensionality, i.e. the large number of variables, is to use Principal Component Analysis (PCA), as explained in subsection 3.2.1. An alternative is to employ a variable selection method, as discussed in subsection 3.2.2. This method selects a subset of the original variables to use in the forecasting models. This can also be implemented together with PCA, where first a subset of the variables is selected and afterward PCA is performed on this subset.

3.2.1 | Principal Component Analysis

One option to reduce the dimensionality, i.e. the large number of variables, is to use Principal Component Analysis (PCA), as explained by Jolliffe (2002). If the variables are correlated with each other, PCA generates components that are linearly uncorrelated, i.e. orthogonal, to each other and explain as much of the variability of the original data as possible. The original variables are reduced to a smaller number of components that are linear combinations of the original variables. These components can subsequently be used as explanatory variables in our model. One disadvantage of doing this is that the interpretability of our model may be lost. The components, themselves, may not have any clear meaning, which means we can not interpret the results of the model, but solely use it for forecasting purposes.

3.2.2 | Filtering Procedure

In this thesis, a correlation based variable selection criterion, proposed by Hall (1999), is used to select an optimal variable subset. This criterion is defined as follows:

$$SubsetScore = \frac{\sum_{i=1}^{K} corr(y, f_i)}{\sqrt{K + (K-1)\sum_{l=1}^{K}\sum_{j=l}^{K} corr(f_l, f_j)}}, \tag{3.2}$$

where K is the number of variables in the subset and $corr(X_1, X_2)$ is the correlation between variable X_1 and X_2. In this case, y is GDP growth and f_i is one of the generated sentiment variables. A greedy backward elimination procedure, as presented by Hall, is used to select a subset. First, the full set of variables is evaluated with Eq. 3.2. After this, a variable is temporarily removed and the truncated set of variables is evaluated with Eq. 3.2. If the truncated set scores higher than the set before, then the variable is permanently removed. If this is not the case, the variable is reinserted in the set of variables. This procedure is continued until all of the variables have been removed once to evaluate the effect of the removal of each variable.

Implementing the formula directly is quite costly in terms of redundant calculations. To illustrate this, Eq. 3.2 is rewritten to the following:

$$SubsetScore = \frac{Relevancy}{\sqrt{K + (K-1)Redundancy}} \tag{3.3}$$

Where $Relevancy = \sum_{i=1}^{K} corr_{y,f_i}$ and $Redundancy = \sum_{l=1}^{K}\sum_{j=l}^{K} corr_{f_l,f_j}$.

Following the greedy backward elimination procedure, the full set of variables is evaluated using Eq. 3.2. Then, a single variable is removed and the truncated set of variables is again evaluated with Eq. 3.2. For each evaluation of Eq. 3.2, a total of $\frac{1}{2}K(K+1)$ correlations are calculated to arrive at the *Redundancy* and a total of K correlations are calculated to arrive at the *Relevancy*. Together, this means that $\frac{1}{2}K^2 + 1\frac{1}{2}K$ correlations are calculated for each set evaluation.

In the worst case scenario, every variable is kept in the filtered set of variables. This results in a total of $K(\frac{1}{2}K^2+1\frac{1}{2}K) = \frac{1}{2}K^3+1\frac{1}{2}K^2$ correlation calculations. This gives a complexity of $\mathcal{O}(K^3 + K^2)$. Nevertheless, a large part of these correlation calculations are redundant. Let *Relevancy* of the initial set of variables be denoted by Rel_{init}, the *Redundancy* of the initial set of variables be denoted by Red_{init}, and the variable that is removed is the K-th variable. The subset score can then also be calculated according to Eq. 3.4.

$$SubsetScore = \frac{Rel_{init} - corr_{y,f_K}}{\sqrt{N + (N-1)(Red_{init} - \sum_{i=1}^{K} corr_{f_i,f_K}}}, \tag{3.4}$$

where $N = K - 1$. This means that each subsequent score calculation, except the first score calculation of the full set of variables, only needs $1+K$ correlation calculations. In the worst case scenario, again, all of the variables are kept, but this time that leads to $K(1+K) + \frac{1}{2}K^2 + 1\frac{1}{2}K = 1\frac{1}{2}K^2 + 2\frac{1}{2}K$ correlation calculations, which results in a complexity of $\mathcal{O}(K^2 + K)$. Obviously, reducing the complexity of the original procedure, as described in Hall, from $\mathcal{O}(K^3+K^2)$ to $\mathcal{O}(K^2 + K)$ leads to a large improvement in computation speed. In this thesis, the computation time was 145 times faster than the original procedure

by using the adjusted equation, a modified correlation calculation formula[4], and a data structure to store the calculated correlations.

3.3 | Forecasting Models

In this section, the different models that are used to forecast GDP are explained. The baseline model, EICIE, is described in subsection 3.3.1. In addition to EICIE both a neural network and a random forest are used. In general, machine learning methods are designed to be able to handle a large amount of variables with a small sample (i.e. $p \gg n$ problems). In section 3.3.2, neural networks are explained and in section 3.3.3, random forests are described.

3.3.1 | EICIE

The baseline model is the EICIE model, as developed by de Groot and Franses (2006). It is an error correction model with a moving average term of 4 lags. The model is defined as follows:

$$log(GDP_t) - log(GDP_{t-4}) = \mu + \beta_1(log(Randstad_t) - log(Randstad_{t-4})) + \beta_2(log(GDP_{t-4} - log(Randstad_{t-4})) + \epsilon_t - \theta\epsilon_{t-4} \quad (3.5)$$

Where the dependent variable is the yearly growth rate of GDP, Randstad is staffing data from a temp agency in the Netherlands and $log(GDP_{t-4} -$

[4]See Appendix A.1

$log(Randstad_{t-4})$ is the error correction term. Finally, $\epsilon_t - \theta\epsilon_{t-4}$ is the moving average term with a lag of 4 quarters.

The resulting components are added to this model with a forward elimination variable selection procedure. This means that each sentiment component is added to the model and only kept if it remains significant. Furthermore, if after the addition of a certain sentiment component, a different sentiment component becomes insignificant, it is removed. The disadvantage to this is that a maximum of around 40 variables can be added, dependent on the sample size. Initially, around 4000 sentiment variables are generated, dependent on the syntactic path restrictions. Only using 40 variables out of 4000 possible variables leads to a large information loss. As previously mentioned, this is why PCA is applied to the raw sentiment variables to create new sentiment components. These are added to the EICIE model. An alternative solution is to apply machine learning models, as these are specifically designed for problems where the number of variables is much higher than the sample size.

3.3.2 | Neural Network

A neural network in its most basic form consists of an input layer (the variables), a number of hidden layers of nodes, and an output layer (the dependent variable), as is illustrated in figure 3.1 and explained in Kosko (1992).

Each node in the first hidden layer is connected to each variable and each node in the next hidden layer or the output layer. The second layer is connected to

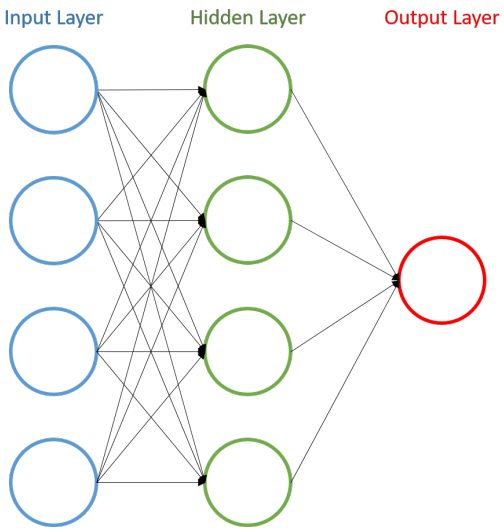


Figure 3.1: A representation of a neural network with 4 input variables(blue), 1 hidden layer(green), and a single output variable(red)

all of the nodes of the first layer and all of the nodes of the next layer or the output layer and so on. More complex neural networks are also possible, as illustrated in figure 3.2, which has three hidden layers.

The connections between the nodes are weights, which are updated during the learning process. The updating of the weights can differ between algorithms, but each time the network is confronted with a new observation of variables, the weights are updated according to some scheme dependent on the learning algorithm. In this thesis, the globally convergent version of the resilient backpropagation, developed by Anastasiadis et al. (2005), is used as the learning method to update the weights. This is a learning algorithm based on resilient backpropagation, which should converge to at least a local optimum.

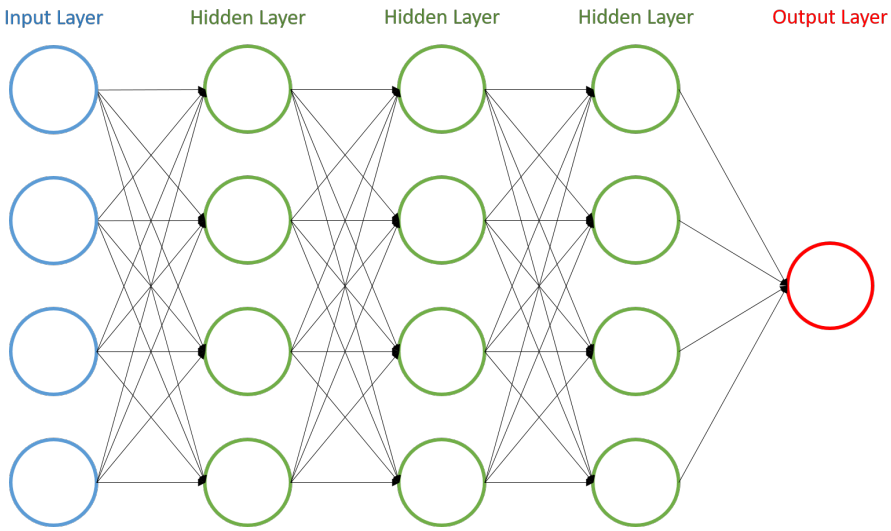


Figure 3.2: A representation of a neural network with 4 input variables(blue), 3 hidden layer(green), and a single output variable(red)

3.3.3 | Random Forest

A random forest is an ensemble method developed by Breiman (2001), where the predictions of multiple individual tree classifiers are combined to a single prediction. In general, a regression tree is considered to be prone to overfitting but has a number of desirable characteristics. Decision trees are invariant to different scaling between the variables and robust to redundant variables. By combining multiple decision trees the desirable characteristics are kept, but the accuracy is improved compared to only a single decision tree.

In this thesis, the implementation of Breiman is used. First, the training sample is randomly sampled with replacement for each tree. This means that each tree is trained on different subsets of the original data, each containing approximately 66% of the full data set. This also creates an out-of-bag part of

the data for each tree (the remaining 34%), which can be used to evaluate the generalization error. Breiman found that the random selection of the training set in combination with random selection of the variables improves the results. This means that a random set of the variables is selected that is considered to be used for a split at each of the nodes. The trees are grown unpruned and according to the CART methodology, as developed by Breiman et al. (1984).

Every tree has a number of nodes and at every node, the variable is selected that gives the best split. This means that if a certain variable is redundant, it will (almost) never be selected at any of the nodes and hence it is virtually ignored in the model. This means that random forests can be seen as a variable selection method. The previously mentioned correlation based method belong to the so-called filter variable selection category, as devised by Chandrashekar and Sahin (2014). Random forests can be seen as so-called embedded variable selection methods. Rather than choosing an optimal subset apriori, i.e. irrespective of the model, the variable selection is part of the construction of random forests.

3.4 | Parameter Optimization

There are a number of possible parameters that needs to be optimized. First, in subsection 3.4.1 the restrictions on the syntactic paths are explained. The machine learning models also have a number of parameters that need to be optimized, as explained in subsection 3.4.2

3.4.1 | Syntactic Path Restrictions

There are 11 syntactic relationships that are used in this thesis, as can be seen in table 3.1. All of these syntactic relationships can be used to restrict the opinions that are used to generate the data. Each syntactic relationship can be either allowed or disallowed. As previously mentioned, each opinion has a syntactic path between the aspect and the sentiment word. If for example, the "mod" relationship is disallowed, then all of the opinions that have a syntactic path with the "mod" relationship are discarded when producing the data. Furthermore, the total length of the syntactic path can be restricted. In this thesis, the maximum length ranges from 1 to 15. In combination with a total of $2^{11} = 2048$ possible combinations of allowed and disallowed syntactic relationships, this gives a total of $15 * 2^{11} = 30720$ possible combinations of parameters. It is possible to evaluate every possible combination and hence a "brute force" approach is used. Every possible combination is evaluated by using a validation sample. The models are estimated on the sample from 2000 till 2007 and the sample from 2008 till 2009 is used as the validation sample. Theoretically, this should mean that the optimal parameters are found if the validation set is a good representation of the test set. Experiments showed that this was indeed the case in this thesis.

3.4.2 | Machine Learning Parameter Tuning

The two machine learning models that are used in this thesis also have a number of parameters that need to be tuned. Tuning is the optimization of model parameters, such as neural networks. Tuning is an important step in model building since these parameters can have a substantial influence on the predictive performance.

Neural Network

The neural network uses layers of nodes, where both the number of layers and number of nodes influence the performance. It was shown by Brown and Harris (1994) that a neural network with three hidden layers can approximate any nonlinear continuous function. Hence, the number of layers is limited to a maximum of three in this thesis. The number of nodes ranges between 0 and the number of variables, where each subsequent layer can not have more nodes than the layer before. This means that the number of nodes in the first layer ranges between 1 and the number of variables (at least one layer has to have one node, otherwise no predictions can be made), the number of nodes in the second layer ranges from 0 to the number of nodes in the first layer, and the number of nodes in the third layer ranges from 0 to the number of nodes in the second layer. These are the only parameters that are tuned for the neural network. Additionally, a variable selection procedure is also employed. Each variable is added to the neural network and if the performance improves, this

variable is kept in the data set. The tuning is done with a grid search of all the possible values of the parameters. For each variable that is added to the data, the grid search is performed to find the optimal number of nodes and the optimal number of layers.

It should be noted that the starting weights of the neural network are randomly initialized and this leads to quite some fluctuations in the results. In theory, these weights should be optimized to a globally convergent optimum, which means the starting weights should not have any influence on performance whatsoever, however, in this thesis this is not the case. This is probably due to the fact that the number of observations is small in comparison to the number of variables. To solve this instability in the parameter tuning procedure, the neural network is estimated 100 times for each configuration of parameters and the best performance is used as the evaluation score. Unfortunately, this also gives rise to one limitation in the parameter tuning procedure. It is only feasible to do this in the case that PCA is applied, which limits the variables to around 65. The number of possible combinations of layers and nodes, as well as the length of the variable selection procedure, is dependent on the number of variables. In the cases where PCA is not applied, the computation time to tune the parameters will be too high and hence a neural network is only trained once for each parameter configuration if PCA is not applied.

Random Forest

The random forest has a number of parameters that need to be tuned. First, the number of variables that are randomly selected for each node in the trees. According to Breiman, this is the only parameter that has an effect on the performance of the random forest. Despite this statement, a number of other parameters are also tuned. The node size, which is the minimum size of the terminal nodes. This means, that if a node has N number of observations and $N < nodesize$, then the node will not be split any further. Setting this parameter higher leads to pruned trees and a shorter computation time. The third parameter is whether to sample with or without replacement. As previously explained, the random forest randomly samples a subset of the original training sample. If this is sampled with replacement approximately 34% will be used as the out-of-bag sample to estimate the generalization error. If this is sampled without replacement, an additional parameter is needed, which is the sample size. This is the number of observations that are used to grow the trees. This means that setting this higher will lead to a smaller out-of-bag sample and vice versa. The out-of-bag sample is randomly selected, but all of the variables are time series. Normally, the out-of-bag sample is used to estimate the generalization error, but this does not make sense if the out-of-bag sample consists of random observations. This is the reason that the sampling size was added to the parameter tuning. These parameters are also tuned by using a grid search. For the data sets where PCA is applied, the grid search is done exhaustively (i.e., using all possible combinations). For the other data

sets, an exhaustive grid search takes too long. For the data set, where only the filtering procedure is applied, the grid search is done by using steps of 2 for the sample size and the node size. For example, the node size starts at 1 and is iteratively increased with 2, i.e., 1, 3, 5, 7, etc.. The number of randomly selected variables is increased with steps of 5, i.e., 1, 6, 11, 16, etc.. For the untransformed data set, the node size and the sample size are again increased with steps of 2, but the number of variables randomly selected is increased in steps of 10. In table 3.3, the maximum value of the different parameters and the step size is summarized, dependent on the data set that is used.

Table 3.3: An overview of the maximum value and the step size of each parameter in the tuning procedure. Var Size is the number of randomly selected variables.

	Node Size		Var Size		With	Sample Size	
	Max	Step	Max	Step	Replacement	Max	Step
PCA	32	1	66	1	FALSE	32	1
Filtered + PCA	32	1	66	1	FALSE	32	1
Filtered	32	2	229	5	FALSE	32	2
Untransformed	32	2	1837	10	FALSE	32	2
PCA	32	1	66	1	TRUE	-	-
Filtered + PCA	32	1	66	1	TRUE	-	-
Filtered	32	2	229	5	TRUE	-	-
Untransformed	32	2	1837	10	TRUE	-	-

3.5 | Evaluation Procedure

In this thesis, all of the models are evaluated by their mean absolute error (MAE). SN publishes the GDP in the Netherlands and the final definitive values of GDP are used as the "real" values of GDP. In some of the last quarters

of 2015, these final definitive values are not available and the most up to date GDP values are used as the real values of GDP. The absolute value of the difference between the forecast of a model and the real value, i.e. the error, is calculated for each forecasted quarter and the average of these is the MAE. In other words, the MAE is calculated according to Eq. 3.6:

$$MAE = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N} \tag{3.6}$$

Where $\hat{y}_i$ is the predicted value of the yearly difference of $\log GDP$ and y_i is the real value of the yearly difference of $\log GDP$. All of the models predict the yearly difference of $\log GDP$, which means that the $MAE \times 100$ is the error in prediction of the yearly growth rate of GDP. In other words, the reported MAE is multiplied by 100 to represent the average deviation of the forecast from the real yearly growth rate of GDP.

In order to evaluate the different models, the sample is split into three parts, a training sample, a validation sample, and a test sample. A number of parameters have to be optimized and for this purpose, the validation set is used. All of the models are only estimated on the training sample and the MAE of the forecasts for the validation sample is used to measure the performance of a certain configuration of parameters. After the optimal parameters are found, the models are estimated on both the training sample and the validation sample. The test sample is used to measure the out-of-sample forecast performance of the models, again using the MAE of the forecasts.

The models are estimated with two procedures to forecast the test sample, a "static" procedure and a "dynamic" procedure. In both procedures, the model is estimated on both the training sample and the validation sample and subsequently used to produce forecasts for the test sample. Each model forecasts $\hat{y}_{t+1}$ using the data available at time t. The difference between the static procedure compared to using the dynamic procedure, is that the model is not re-estimated after each forecast. In the dynamic procedure, however, the models are re-estimated after each forecast. For example, first a random forest is estimated on the sample 2000Q1 to 2009Q4 and this model is used to forecast 2010Q1. After this, the random forest is re-estimated on the sample 2000Q1 to 2010Q1 and subsequently used to forecast 2010Q2 and so on.

In order to determine if certain forecasts are significantly more accurate than others, statistical test of different population means are utilized. Diebold and Mariano (2012) developed a test statistic for this purpose, but in case only 1-step ahead forecasts are made (as is the case in this thesis), their test statistic is equivalent to the paired t-test statistic. Let $d_t = g(e_{1,t}) - g(e_{2,t})$ be the loss-differential, where g is the error function, in this case $g(x) = |x|$. Testing for a significant difference in forecasting accuracy can now be done by testing the null hypothesis that the population mean of the loss-differential is 0, i.e., $H_0 : E[d_t] = 0$. The paired t-test uses $t = \frac{\bar{d}-\mu_0}{s_d/\sqrt{n}}$ to test the null hypothesis that the population mean is equal to μ_0, 0 in this case.

A critical assumption of a paired t-test is that the loss-differential follows a normal distribution. In order to test this, the Jarque-Bera test, developed by

Jarque and Bera (1987), is used to test for normality. Jarque and Bera defines the following statistic:

$$JB = \frac{n-k+1}{6}(S^2 + \frac{1}{4}(C-3)^2) \tag{3.7}$$

Where n is the number of observations, k is the number of regressors or 1 in case a single time series is tested. S is the sample skewness, as defined in equation 3.8

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{\frac{3}{2}}} \tag{3.8}$$

C is the sample Kurtosis, as defined in equation 3.9

$$C = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \tag{3.9}$$

Where x_i is the observation of the tested time series at $t = i$ and $\bar{x}$ is the sample mean of x. This test statistic follows a Chi-square distribution with 2 degrees of freedom under a null hypothesis of a normal distribution. This means that normality can safely be assumed in case the null hypothesis of the Jarque-Bera test is not rejected.

Diebold and Mariano point out that it is useful to use non-parametric tests to complement their test statistic in situations with only a few forecast error observations. In this thesis, only 24 forecasts have been made and hence the Wilcoxon signed-rank test, as described by Woolson (2008), is used to complement the results of the paired t-test. Moreover, the Wilcoxon signed-rank test relaxes the assumption of normality and only assumes a symmetric

distribution. If the loss-differential is sorted according to size, then ranks can be assigned from 1 (the smallest loss-differential) and upwards. The test statistic is subsequently defined as $W = \sum_{t=1}^{T} sign(d_t) \times Rank(|d_t|)$, also called the sum of signed ranks. This test statistic follows a specific distribution with a mean of 0 and a variance of $\frac{T(T+1)(2T+1)}{6}$. In order to calculate a p-value corresponding to a certain value of W, $\frac{W}{\frac{T(T+1)(2T+1)}{6}}$, which follows a standard normal distribution, is used.

4 | Results

In this chapter, the results are presented and discussed. In section 4.1, the effect of the restriction on the syntactic paths are evaluated and discussed. In section 4.2, the effects of filtering and applying PCA are evaluated and in section 4.3, the effect of adding the sentiment variables is evaluated. Finally, in section 4.4, the performance of the most accurate models is discussed and compared to the official flash-estimates of SN and the published estimates of the EICIE model.

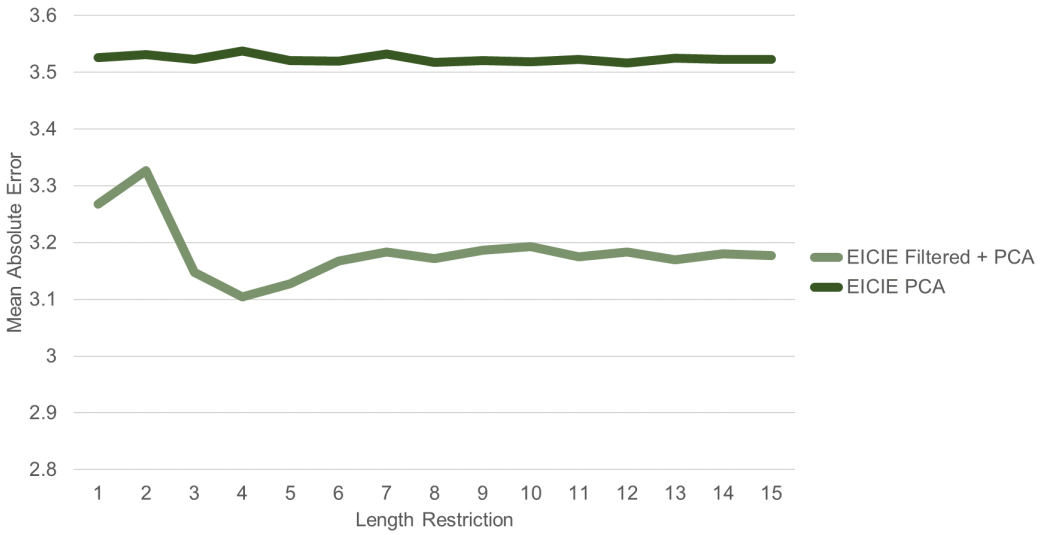
4.1 | Sentiment Parameters Effect on Performance

As previously mentioned, the syntactic path restrictions were tested on the validation set for each possible combination. A total of 11 syntactic relationships could be either allowed or disallowed and the total length of the syntactic path ranges between 1 and 15. In this section, the effect of these restrictions is evaluated. First, the maximum length parameter is evaluated with the EICIE model in subsection 4.1.1. Second, the effect of the syntactic path restrictions is evaluated in subsection 4.1.2. This is evaluated with only the EICIE model in combination with the components from PCA of the filtered and unfiltered data set since this model performs best for the validation sample.

4.1.1 | Length Effect

In order to evaluate the impact of the maximum length parameter, the MAE that was found with EICIE is used. In figure 4.1 the results for the EICIE models are displayed. The EICIE model can only contain a limited number of regressors and hence results can only be obtained if PCA is applied, either on the filtered data set or the untransformed data set.

Figure 4.1: The MAE for each possible value of the length parameter using the EICIE model. This is the average MAE for all possible syntactic path restrictions.



As can be seen in figure 4.1, the influence of the maximum length parameter is strongest if it is very restrictive. For a length of 6 to 15, the performance is quite stable at around 3.20. If it is set to 4, the MAE is at its lowest at around 3.1. For a maximum length lower than 3, the impact is more prominent

with a MAE of around 3.35 at its worst. Note that the MAE in figure 4.1 is the average MAE over all the possible syntactic path restrictions. The best MAE was found with a length set at 10. This result is not counter-intuitive since this is not a very restrictive length (there are not many opinions that are discarded), but it does filter out the noisiest opinions. Note that by applying a very restrictive combination of syntactic path restrictions with the length set at 10; this could still lead to a data set with many discarded opinions.

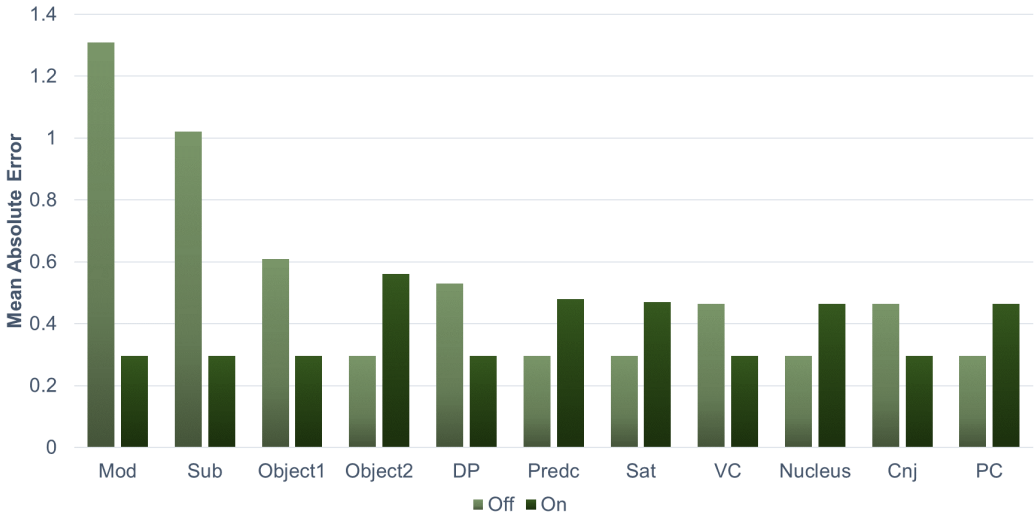
Furthermore, it can be seen that the effect of the length parameter is quite low for the EICIE model that uses the PCA components of the untransformed data set. It should be noted that the untransformed data set is quite sparse with many aspects that were not mentioned in a certain time period and hence with scores equal to 0. The filtering procedure discards these aspects since the correlation with GDP growth will also be low in these cases. This is probably the reason that the filtering improves performance and that the performance without filtering is insensitive to the length parameter.

4.1.2 | Syntactic Path Restrictions Effect

In figure 4.2, the best MAE is plotted for each syntactic relationship, either it is allowed(On) or disallowed(Off). The reported MAE's in figure 4.2 are achieved by using the EICIE model with the PCA components of the filtered data set. Additionally, the reported MAE is the lowest MAE achieved for all parameter configurations, while a certain syntactic path restriction is allowed or disallowed. For example, a MAE of around 0.3 is obtained, when the "Mod"

relationship is allowed, but the lowest MAE that is achieved, while the "Mod" relationship is disallowed, is around 1.3.

Figure 4.2: The lowest MAE achieved for all possible parameter configurations, while a certain syntactic relationships were either allowed or disallowed.



It can be seen in figure 4.2 that both "Mod" and "Sub" are the most influential syntactic relationships. This seems intuitive since each subject is a noun and hence more likely to be an aspect. Additionally, the modifier relationship most frequently occurs in unison with an adjective, which is probably a sentiment word. If either of these is disallowed, then there are probably many opinions that are discarded from the data. This large information loss is most likely to be the reason for the large discrepancy between being allowed and disallowed for these two syntactic relationships. The third largest difference in performance between being allowed or disallowed is for the "Obj1" relationship. This is most likely due to the fact that a sentiment about a noun in the object

is propagated through to the subject in the sentence if this syntactic relationship is allowed. For example, in the sentence: “The iPhone has a great screen”, the positive sentiment from “great” is also transmitted to “iPhone” through the object and subject relationship. A counter-intuitive result is that the “Predc” relationship decreases the performance if it is allowed. This relationship occurs in sentences, such as: “The iPhone is beautiful”, where the sentiment from “beautiful” directly affects the subject in the sentence. However, the “Predc” relationship also occurs when the adjective affects the object in the sentence. For example, for a sentence, such as “He views the iPhone as beautiful”, then the sentiment from “beautiful” is also propagated through to the subject in the sentence, i.e., “He” in this case. It could be that this produces more noise than useful information in the sentiment data and consequently decrease the predictive accuracy. There are not any other noteworthy performance gains or losses with the remaining syntactic relationships. The optimal configuration is with Mod, Sub, Obj1, DP, VC, and Cnj allowed and the rest disallowed. This configuration is used throughout the rest of this section.

4.2 | Effect of Filtering & PCA

All of the models are evaluated with variables from both the unfiltered and the filtered data set, as well as with components from the PCA procedure applied to these two data sets. This means that for each model four different data sets were used, except for the EICIE model, where only components from the PCA could be added to the model. The results of applying PCA are shown in

table 4.1. Here, the improvement in the MAE due to the application of PCA is shown. This means that the difference between the performance of a model without PCA and a model with PCA is shown. Hence, positive numbers is an improvement in the MAE and negative numbers mean that applying PCA gives worse results.

Table 4.1: Decrease of the MAE due to the application of PCA to the data sets.

Model	**Static**	**Dynamic**
Neural Network	1.53	0.073
Neural Network Filter	1.275	1.56
Random Forest	0.004	0.021
Random Forest Filter	0.551	0.497

As can be seen in table 4.1, applying PCA gives an improvement for every model, albeit small in some cases. Applying PCA to the unfiltered data set with a random forest only gives negligible improvements. Applying PCA to the filtered data set with a random forest gives an improvement in the MAE of 0.55 and 0.50 for the static and dynamic forecasting, respectively. The highest performance gain due to the application of PCA occurs when PCA is applied to the filtered data set and when a neural network is used as the forecasting model. Here, the application of PCA gives an improvement of 1.28 and 1.56 for the static and dynamic forecasting, respectively. When PCA is applied to the unfiltered data set with a neural network as the forecasting model, the improvement is contradictory. With the static forecasting, the improvement is 1.53, the biggest improvement when using the static forecasting procedure, but only 0.07 when using the dynamic forecasting. Recall from section 3.4 that

there are some instability issues with the neural network. The starting weights have a pronounced influence on performance, which is solved in the cases when PCA is applied, which limits the number of variables, but not in the cases where PCA is not applied. This is probably the reason for these contradictory results. The application of PCA has a high impact on the forecasting performance. In all cases, applying PCA leads to a varying increase in performance. On average, applying PCA leads to a 38.8% improvement of the forecasting accuracy.

The results of applying the filtering procedure are shown in table 4.2. Here, the improvement in the MAE due to the application of the filtering procedure is shown. This means that the difference between the performance of a model without the filtering procedure and a model with the filtering procedure is shown. Hence, positive numbers is an improvement in the MAE and negative numbers mean that applying the filtering procedure gives worse results.

Table 4.2: Decrease of the MAE due to applying the filtering procedure to the data sets.

Model	**Static**	**Dynamic**
Neural Network	0.835	-0.753
Neural Network PCA	0.58	0.734
Random Forest	0.034	0.133
Random Forest PCA	0.581	0.609
EICIE PCA	1.28	0.757

As can be seen in table 4.2, the filtering has a substantial impact on the predictive accuracy. For the neural network the effect of filtering, but without applying PCA, is pronounced, but conflicting. When using the static forecasting procedure, filtering decreases the MAE with 0.84, however, when using the

dynamic forecasting procedure, filtering increases the MAE by 0.75. Recall from section 3.4 that there are some instability issues with the neural network. The starting weights have a pronounced influence on performance, which is solved in the cases when PCA is applied, which limits the number of variables, but not in the cases where PCA is not applied. This is probably the reason for these contradictory results. In the case that PCA is applied, it can be seen that filtering gives a performance boost of 0.58 and 0.73, using the static and the dynamic forecasting procedure, respectively. This is a promising result, as the filtering procedure more than doubles the forecasting accuracy.

For the random forest, the effect of applying the filtering procedure is equally promising and consistent. In the case that PCA is not applied, the improvement of the performance due to filtering is 0.03 and 0.13 for the static and dynamic forecasting, respectively. This increase is quite low compared to the case when PCA is applied, where the improvement of the performance is 0.58 and 0.61 for the static and dynamic forecasting, respectively. Again, using the filtering procedure doubles the forecasting accuracy.

For the EICIE model, the improvement of the forecasting accuracy is the highest compared to the rest of the models. The improvement is 1.28 and 0.76 for the static and dynamic forecasting procedures, respectively. It can safely be said that the filtering procedure has shown itself to have a big impact on performance. On average, filtering leads to 32.5% improvement in forecasting accuracy over all the models. In the case that only the results where PCA is applied are taken into account the filtering leads to a 56.2% improvement in

forecasting accuracy.

The PCA components of the filtered data set give the best performance for every model and henceforth this data set will be used. Furthermore, the random forest uses a node size of 17 and the number of randomly selected variables is set to 63. The sampling is done with replacement. The neural network uses 3 layers with 4, 4, and 3 nodes for the first, second, and third layer, respectively. These parameter settings stem from the parameter tuning procedure when using the PCA components of the filtered data set.

4.3 | Effect of the Sentiment Variables

In this section, the effect of adding the generated sentiment variables is discussed. For all of the models, the PCA components of the filtered data set are used, since these give the best results. In table 4.3, the decrease of the MAE due to the addition of the sentiment variables is displayed with both the static and the dynamic forecasting procedure.

Table 4.3: Decrease of the MAE due to the addition of the Sentiment variables.

Model	**Static**	**Dynamic**
EICIE Filter+PCA	2.104	0.729
Random Forest Filter+PCA	0.703	0.419
Neural Network Filter+PCA	0.655	0.534

As can be seen in table 4.3, using only the Randstad staffing data gives the worst performance for all of the models, both with static and dynamic fore-

casting. The addition of the sentiment variables improves the forecasting performance quite significantly. As can be seen, the decrease of the MAE is 2.104 at the highest and 0.419 at the lowest.

As previously mentioned, in order to test for a significant difference in forecasting accuracy between models, 2 tests are employed, i.e. the Wilcoxon signed-rank test (WSR test) and the paired t-test (Paired t-test). These tests are used to test for a significant improvement in prediction accuracy due to the addition of the sentiment variables. In table 4.4, an overview of the results of the tests are given.

Table 4.4: P-values for the different tests of significant improvement in forecasting accuracy due to the addition of sentiment variables.

Model	**Paired t-test**	**WSR test**
EICIE Static	0.000	0.000
EICIE Dynamic	0.002	0.003
Random Forest Static	0.004	0.030
Random Forest Dynamic	0.021	0.038
Neural Network Static	0.003	0.009
Neural Network Dynamic	0.011	0.049

As can be seen in table 4.4, the addition of the sentiment variables give a significant improvement in forecasting accuracy for all of the models according to both the Wilcoxon signed-rank test and the paired t-test at a 5% confidence level. All of the loss-differentials are tested for normality using the Jarque-Bera test and are all shown to follow a normal distribution at a 5% confidence level. This means that the assumption of the paired t-test is valid. Moreover, the Wilcoxon signed-rank test supports all of the outcomes of the paired t-test. It can safely be concluded that the addition of the sentiment variables that

are generated by the framework developed in this thesis lead to a significant improvement in the prediction accuracy for all of the models.

4.4 | Comparison with SN Flash-Estimates & the Published EICIE Estimates

In this section, an overview of the performance of the models with the best data sets and variables are given. Furthermore, a comparison is made with the SN Flash-Estimates and the published EICIE estimates that were adjusted according to expert opinions (EICIE+Expert Opinions). The forecasts are made both with only the sentiment variables (denoted as SV) and both the sentiment variables and the Randstad staffing data combined (denoted as SV+RS). In table 4.5, the MAE of the different models are displayed, both with static and dynamic forecasting. All of the models use the PCA components of the filtered data set.

As can be seen in table 4.5, leaving the Randstad staffing data out of the models does not affect the results much. For the EICIE model, it actually improves the performance slightly, but only with 0.03 at the highest. For the random forest, the performance drops slightly, but again only with 0.035. For the neural network, it does not affect performance at all, but this is due to the variable selection procedure that is done during the parameter tuning. The Randstad staffing data is left out by this procedure in any case.

Table 4.5: Overview of the MAE of the different models with only sentiment variables (SV), as well as with both the sentiment variables and the Randstad staffing data (SV+RS).

Model	**Static**	**Dynamic**
EICIE SV	0.582	0.648
EICIE SV+RS	0.598	0.678
EICIE RS	2.702	1.407
Random Forest SV	0.618	0.572
Random Forest SV+RS	0.583	0.558
Neural Network SV	0.497	0.486
Neural Network SV+RS	0.497	0.486
EICIE+Expert Opinions	-	0.929
SN Flash-Estimates	-	0.320

The best performance overall is achieved by using a neural network with the PCA components of the unfiltered data set and without the Randstad staffing data. The dynamic forecasting procedure gives the best results, but it is only a small improvement compared to the static forecasts. This is also expected, as the model is re-estimated after each forecast of an individual quarter in the sample. The random forest and the EICIE model perform quite similar with static forecasts, but there is a difference in favor of the random forest when using the dynamic forecasting procedure.

The worst performance is achieved by using the EICIE model with only the Randstad variables. The published EICIE estimates, however, are adjusted according to expert opinions and result in a MAE of 0.929, which is considerably better. The previously mentioned tests are employed to test for a significant improvement in forecasting accuracy compared to the published EICIE estimates. In table 4.6, an overview of the results of the significance tests are

displayed.

Table 4.6: P-values for the different tests of significantly lower forecasting errors by the published EICIE estimates compared to the forecasts of the other models.

Model	**Paired t-test**	**WSR test**
EICIE SV Static	0.013	0.023
EICIE SV Dynamic	0.034	0.055
Random Forest SV+RS Static	0.016	0.023
Random Forest SV+RS Dynamic	0.012	0.011
Neural Network SV Static	0.003	0.005
Neural Network SV Dynamic	0.002	0.005

As can be seen in table 4.6, all of the models are significantly more accurate compared to the published EICIE estimates, except for the EICIE model with only sentiment variables and with dynamic forecasting. For this model, only the paired t-test rejects the null hypothesis at a 5% confidence level, while the Wilcoxon signed-rank test gives a p-value of 5.5%. All of the loss-differential are again tested for normality using the Jarque-Bera test and the null hypothesis can not be rejected for any of the loss-differentials. It can safely be assumed that the loss-differentials follow a normal distribution and hence the paired t-test is considered to be valid in this case. Moreover, since the Wilcoxon signed-rank test supports the paired t-test in almost every case it can safely be concluded that the models with the sentiment variables are significantly more accurate than the published EICIE estimates.

Moreover, the SN publishes their flash-estimate 45 days after the end of a quarter. The SN flash-estimates are the best in terms of forecasting accuracy with a MAE of 0.32. Ideally, the performance would be matched by our models.

However, this is not the case. To test for a significantly higher forecasting accuracy of the SN flash-estimates compared to the models developed in this thesis, the aforementioned tests are employed again. In table 4.7, the outcomes of the tests are displayed.

Table 4.7: P-values for the different tests of significantly lower forecasting errors by the SN flash-estimates compared to the forecasts of the other models.

Model	**Paired t-test**	**WSR test**
EICIE SV Static	0.018	0.036
EICIE SV Dynamic	0.009	0.041
Random Forest SV+RS Static	0.004	0.012
Random Forest SV+RS Dynamic	0.013	0.081
Neural Network SV Static	0.041	0.115
Neural Network SV Dynamic	0.062	0.133

As can be seen in table 4.7, the results are more contradictory than before. The paired t-test rejects the null hypothesis for all of the models with a confidence level of 5%, except for the neural network with dynamic forecasting, where the null hypothesis is rejected only at a 10% confidence level. The Wilcoxon signed-rank test rejects the null hypothesis of equal forecasting accuracy with a 5% confidence level for both the EICIE models and for the random forest with both sentiment variables and Randstad staffing data and with static forecasting. For the random forest with both sentiment variables and Randstad staffing data and with dynamic forecasting the null hypothesis is also rejected, but only with a 10% confidence level. Since both the paired t-test and the Wilcoxon signed-rank test do not reject the null hypothesis of equal forecasting accuracy for the neural network with only sentiment variables and with dynamic forecasting, it is not possible to conclude that the SN flash-estimates are significantly more

accurate than the forecasts of the neural network.

The comparison is not fair, though, since SN probably uses the same calculation method and variables to calculate the flash-estimate and the final definitive GDP estimates. Since the final definitive GDP estimates are used as the "real" GPD growth. This means that the flash-estimates have an advantage compared to the models developed in this thesis.

All of the loss-differentials are tested for normality using the Jarque-Bera test. The null hypothesis of following a normal distribution can not be rejected for any of the loss-differentials, except for the EICIE model with only sentiment variables and with static forecasting. This means that the paired t-test is valid for every model, except the EICIE model with only the sentiment variables and with static forecasting. However, The Wilcoxon signed-rank test that relaxes the assumption of normality does not give a different result compared to the paired t-test.

5 | Conclusion & Future Work

In this thesis, a sentiment determination framework to extract quantifiable information from news articles has been proposed. The purpose of using news articles to forecast GDP is that there is no publication lag with these online sources and using the determined sentiment variables from the news articles might contain previously untapped information that can improve the forecasting accuracy of existing models. The model made by de Groot and Franses was used as the baseline and the prediction accuracy of EICIE is significantly improved upon with a 65.4% improvement and an 81.6% improvement when using static and dynamic forecasting, respectively. This performance was achieved by using a neural network without the Randstad staffing data, which means that these forecasts have no publication lag at all. A neural network is the least restrictive model of all the models that were used and could theoretically approximate any nonlinear continuous function, which supports its superior performance. Furthermore, some dimensionality reduction and variable selection techniques were shown that have consistently improved the accuracy for all of the models.

The first suggestion for future work is to increase the training sample. The online archive of the "Volkskrant" goes further back than the year 2000 but was not used due to time limitations. Another suggestion is to expand the parameter optimization process. In this thesis, the link restrictions were only

used as either on or off, but these can also be used by their range from 0 to 15 to give a more fine-grained syntactic path restriction. Additionally, a sentiment word list was used of around 3000 sentiment words, however, these might not be the most suited for this purpose. The sentiment words are general and not domain-specific since a domain-specific sentiment word list does not exist in Dutch yet. It could be a useful extension to use the validation sample to filter out sentiment words that introduce noise. Moreover, the score used to generate the sentiment variables did not take the number of opinions into account but measured the sentiment as a relative score. It could be useful to take the number of opinions used to generate the score into account. The final suggestion for future work is to incorporate domain-specific rules. For example, if unemployment increases this is negative, but if consumer spending increases this is positive. These kinds of rules could increase prediction accuracy even further.

A | Appendix

A.1 | Correlation Calculation

The most commonly used formula to calculate the correlation between two variables, say X and Y, is the following:

$$corr = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}} \tag{A.1}$$

This is straightforward to interpret, but harder to compute. The reason is that first, the average of both X and Y has to be calculated. This is already one iteration from 1 to N. Afterward, another iteration from 1 to N is necessary to calculate the mean subtracted values. In practical terms, this means that two iterations from 1 to N are necessary. An alternative option is to use the following formula:

$$corr = \frac{N\sum XY - \sum X \sum Y}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}} \tag{A.2}$$

As can be seen from Eq. A.2, this formula only needs one iteration from 1 to N. This gives a considerable speed-up, especially when calculating millions of calculations.

Bibliography

Anastasiadis, A. D., Magoulas, G. D., and Vrahatis, M. N. (2005). New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, 64:253–270.

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., and Rünstler, G. (2011). Short-Term Forecasts of Euro Area GDP Growth. *The Econometrics Journal*, 14(1):C25–C44.

Baffigi, A., Golinelli, R., Parigi, G., et al. (2002). *Real-Time GDP Forecasting in the Euro Area*, volume 456. Citeseer.

Blinder, A. S. and Zandi, M. (2010). Stimulus worked. *Finance & Development*, 47(4):14–17.

Bouma, G., Van Noord, G., and Malouf, R. (2001). Alpino: Wide-coverage computational analysis of dutch. *Language and Computers*, 37(1):45–59.

Bozic, C. and Seese, D. (2011). News Analytics: Exploring Predictive Power of Aggregated Text Sentiment Measure. In *Proceedings of Annual Paris Conference on Money, Economy and Management*.

Breiman, L. (1996). Bagging Predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Brown, M. and Harris, C. J. (1994). Neurofuzzy adaptive modelling and control.

Callen, T. (2008). What is gross domestic product? *Finance & Development*, 45(4):48–49.

Chandrashekar, G. and Sahin, F. (2014). A Survey on Feature Selection Methods. *Computers & Electrical Engineering*, 40(1):16–28.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(s1):2–9.

Choi, Y. and Cardie, C. (2008). Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics.

Claessens, S. and Ayhan Kose, M. (2009). What is a recession? *Finance & Development*, 46(1):52–52.

de Groot, B. and Franses, P. H. (2005). Real Time Estimates of GDP Growth. *Econometric Institute Report*, (2005-01).

de Groot, B. and Franses, P. H. (2006). Long-Term Forecasts for the Dutch Economy. *Econometric Institute Report*, (2006-06).

De Smedt, T. and Daelemans, W. (2012). "vreselijk mooi!"(terribly beautiful): A subjectivity lexicon for dutch adjectives. In *LREC*, pages 3568–3572.

den Reijer, A. H. (2005). Forecasting Dutch GDP using Large Scale Factor Models. Technical report, Netherlands Central Bank, Research Department.

Diebold, F. X. and Mariano, R. S. (2012). Comparing predictive accuracy. *Journal of Business & Economic Statistics*.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 55(4):665–676.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato.

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kitching, J., Blackburn, R., Smallbone, D., and Dixon, S. (2009). Business strategies and performance during difficult economic conditions.

Kosko, B. (1992). *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence/book and disk*. Prentice Hall, Upper Saddle River.

Kourentzesa, N. and Petropoulosa, F. (2014). Increasing Knowledge Base for Nowcasting GDP by Quantifying the Sentiment about the State of Economy. Technical report, Working Paper.

Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, 72(4):518–550.

Moghaddam, S. and Ester, M. (2010). Opinion Digger: an Unsupervised Opinion Miner from Unstructured Product Reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1825–1828. ACM.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational linguistics*, 37(1):9–27.

Rünstler, G., Barhoumi, K., Benk, S., Cristadoro, R., Den Reijer, A., Jakaitiene, A., Jelonek, P., Rua, A., Ruth, K., and Van Nieuwenhuyze, C. (2009). Short-Term Forecasting of GDP using Large Datasets: a Pseudo Real-Time Forecast Evaluation Exercise. *Journal of forecasting*, 28(7):595–611.

Schumacher, C. and Breitung, J. (2008). Real-Time Forecasting of German

GDP based on a Large Factor Model with Monthly and Quarterly Data. *International Journal of Forecasting*, 24(3):386–398.

Tuckett, D., Ormerod, P., Nyman, R., and Smith, R. E. (2015). Information and Economics: A New Way to think about Expectations and to Improve Economic Prediction. In *Institute for New Economic Thinking 2015 Plenary conference," Liberté, égalité, fragilité," April*, pages 8–11.

Vosen, S. and Schmidt, T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting*, 30(6):565–578.

Vosen, S. and Schmidt, T. (2012). A Monthly Consumption Indicator for Germany Based on Internet Search Query Data. *Applied Economics Letters*, 19(7):683–687.

Woolson, R. (2008). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*.

Zhu, J., Wang, H., Tsou, B. K., and Zhu, M. (2009). Multi-Aspect Opinion Polling from Textual Reviews. In *Proceedings of the 18th ACM conference on Information and Knowledge Management*, pages 1799–1802. ACM.