A NEW APPROACH TO DESIGN TEXT MINING BASED TIME-SERIES FORECASTING SYSTEMS

MASTER THESIS
PETER MICHAEL BAUMGARTL
M.Sc IN OPERATIONS AND SUPPLY CHAIN MANAGEMENT
FACULTY OF ENGINEERING AND MANAGEMENT
AALBORG UNIVERSITY
31 MAY, 2018

AALBORG UNIVERSITY
STUDENT REPORT

School of Engineering and Science
Fibigerstræde 16 DK - 9220 Aalborg East
Phone +45 99 40 93 09
lft@m-tech.aau.dk
www.en.ses.aau.dk

Title:
A new approach to design text mining based time-series forecasting systems

Theme:
Machine Learning and Forecasting

Thesis Period:
01/02/2018 - 01/06/2018

Author:
Peter Michael Baumgartl

Supervisor:
Peter Nielsen

Page Numbers: 53 pages

Date of Completion:
31st of May 2018

Synopsis:

The theme of this master thesis in Operations & Supply Chain Management is machine learning and forecasting in the context of text mining based time-series forecasting.
The overall aim is to identify gaps in the current literature and address them. The initial problem is investigating the state of the research body up to this point. Based on this, holes in literature are identified and their reasons and implications are reviewed. In the problem analysis, it is concluded that the behavioural economic basis of a forecasting system and the previously chosen methods influence the further design process. Since the current framework does not include these findings, in the solution a new approach to design text mining based time-series forecasting systems is developed.

Preface

This is a master thesis on the master's degree programme Operations and Supply Chain Management at Aalborg University. The thesis is written by Peter Michael Baumgartl in the period from February 1^{st} 2018 to June 1^{st} 2018 and the theme for the project is machine learning and forecasting.
Throughout the report, theory used to write this report is explained, and therefore no prerequisites are needed for the reader. Nevertheless, a reading guide will be presented in the following section.
A special thanks to Peter Nielsen, who has been the supervisor for this thesis and was always available to discuss approaches and provide new great ideas. Also thanks to Paul Stockhammer, Andreas Hofbauer and Rasmus Lundgaard Christensen for their comments and thoughts on the thesis. Lastly, a thank you to Anna Paulmichl and my lovely mother and sisters for the emotional support.

Reading guide

The project is structured in three main parts; pre-analysis, problem analysis, and the solution. The pre-analysis is analysing the current research body of text mining based time-series forecasting. The problem analysis provides an in depth investigation of selected topics, which are identified over the course of the pre-analysis. In the last part of the project, the solution presented and evaluated. It is further presenting aspects that require future research attention.

The Harvard method is used for citations in the report, which presents the references as: [Last name of author or name of the source, year]. For example a regular source is referenced as [Peter Nielsen, 2018] or Peter Nielsen [2018], if the author is mentioned directly in the text. The remaining information about the reference can be found in the bibliography, which is listed in alphabetic order. Figures and tables are numbered according to their respective sections.

Summary

In the past couple of years machine learning has made great strives and has found applications in many different fields. Anyhow, one application of machine learning that has received relatively little research attention yet is the combination of text mining and forecasting methods in order to process unstructured data, in other words text, to predict time-series. This interdisciplinary approach, text mining based time-series forecasting, is the theme of this master thesis.
The interdisciplinary distinguishes this forecasting approach from classical machine learning, in that instead of two steps, data collection and machine learning, multiple different methods must be applied to the unstructured input data before it can be processed by a machine learning algorithm. Further, in regular forecasting systems that work with structured data only it is able to mathematically compute the relationship between the variable that is predicted and it's predictors. This is not the case when text data is used as predictor, which requires the introduction of an additional concept to link the input and output of the forecasting system on a logical basis, behavioural economics. The introduction of this new factor, combined with the fact that not that many works have been published in the field so far, have the effect that until now researches have solely focused on how this technique can be used for market prediction.
As mentioned above this field is relatively young, which means there are still some fundamental aspects to it that are not fully understood yet. One of them is the influence that the underlying behavioural economic basis has on the methods that are feasible in a certain forecasting system configuration. Another unknown factor is the influence of the specific problem itself and the methods chosen in certain layers of the system on the overall system design. This thesis shows that the impact that the behavioural economic basis and the decisions made in the feature selection have a severe impact on system design. Based on this finding two different approaches to construct a text mining based time-series forecasting system are identified, one that is based on using the body of the text document itself and one that categorises the text documents and then uses the obtained classes for prediction.
This insight was used to create a new approach to design such forecasting systems that incorporates the effect that the different behavioural economic theories and methods have. To show the application of this new method, the real world forecasting example of forecasting intraday prices was introduced and a text mining based time-series forecasting system was developed to solve it, using micro-blogs published on twitter. This was done testing two different machine learning algorithms as binary classifiers, neural networks and a tree based method, where the latter shows a prediction accuracy of forecasting a up- or down movement of the market that performs better than a random approach.
While the conclusion of the analysis and the solution show how certain decisions in the creation process of a text based forecasting systems influence the feasibility of methods in other system layers, it remains to be seen if certain method combinations perform better than others. Further it is discussed that the complexity of such a system prevents it from being optimised holistically. Future research areas that arise from the thesis are on one hand to investigate if promising method combinations can be found and on the other hand

how this method can be applied in other fields.

Contents

1. Introduction

In today's world, artificial intelligence and machine learning are more and more becoming a bigger part of daily life. From smart speakers that understand speech and act on it, to self-driving cars that are able to interpret and act autonomously with their environment, to Google's duplex that can make phone calls on it's own, seemingly every month a new break through is presented. Although all these exiting new advances are made, the general sense is that in machine learning currently only a fraction of the potential is used. In its core, machine learning's foundation is to find patterns in the past that can be used to predict the future. This responds to what the famous author George Orwell [1949] already claimed in 1949 in his book "1984":

"He, who controls the past controls the future."

In contrast to Orwell, who uses this quote in quite a pessimistic context, machine learning seeks to understand the reality of the past to use this knowledge to be able to anticipate the future.

While both time-series forecasting and the processing and classification of text are some of the core disciplines of machine learning, relatively little research has been done on how text can be used to forecast time-series. This field, text mining based time-series forecasting, is the theme of this thesis. Naturally, humans are very good at processing language. If a lot of people claim *"X is going to happen tomorrow."*, it is easy for a person to make a prediction about X, based on the unstructured data, e.g. the text, they have access to. While humans are able to use unstructured data intuitively and act based on it, computers are used to work with structured data, e.g. numbers. Therefore they need to process the unstructured data first, to translate it into something they can process. This fact makes text mining based time-series forecasting systems more complex than regular machine learning methods on one hand, but on the other hand this special feat is also what makes this field especially interesting to work with.

This thesis starts out with a general introduction to text mining based time-series forecasting. Afterwards, it is shown that the limited amount of research that has been conducted so far leads to some areas within the field that are not very well understood. One of this areas is the influences that certain aspects have on the design of such a forecasting system. This topic is investigated in detail in this thesis and it is concluded that when designing a text mining based time-series forecasting system it is imperative to consider some pre-requisites. Based on this finding, a new approach to design these forecasting systems is presented and it is shown how this approach can be used in practice.

2. The Initiating Problem

This chapter introduces the initial problem description for this thesis, which is based on the general research topic of using text mining techniques for time-series forecasting.

2.1 Initiating description

The emergence of the world wide web and advancements in technology in recent decades have led to a surge in the availability of data. While Google indexed about one million web pages in 1998, in 2008, this value already exceeded one trillion. [Wei Fan, 2013] This has changed the way how information is created and consumed. Social networks like Twitter and Facebook have made it possible for everyone to publish their opinion and make it available worldwide. It is estimated that by 2020 6.6 zettabytes of data will be created, replicated and consumed in the United States alone. [John Gantz, 2013] This development has gone hand in hand with a rising research focus on big data and especially data mining, seeking to identify patterns in a given set of information. Classical data mining methods require a highly structured format of data, which limits it's application to so called structured data, meaning that for every member of a observed sample either ordered numerical or categorical attributes are known. [Sholom M. Weiss, 2010] This is not the case with text data, which is unstructured by definition since there are no special requirements that need to be fulfilled when composing a document, and therefore data mining methods cannot be applied to it. The combination of availability of this form of unstructured data and the lack of methods to process it has led to the emergence of a new research field, text mining. This field focuses mainly on the question of how text documents can be classified autonomously, but there has also been some effort to use unstructured data as a predicting factor for time-series forecasting, which is the main topic of this thesis.

2.2 Initiating problem statement

Above, a general description of the background of the research topic is provided. Based on this, an initiating problem statement is formulated, which will set the grounds for the pre-analysis of the general research topic.

What is the current research body of text mining techniques for time-series forecasting? What are the predominant gaps in literature?

The pre-analysis will investigate the initial problem statement by analysing the available literature on the topic, in which the focus will be on determining gaps that require further research.

3. Methodology

This chapter will provide an explanation of the way this project was approached. First, the thesis structure will be explained, outlining the format of the project and the logical steps that were taken. Following, this chapter provides an overview of the different methods used in each part of the thesis.

3.1 Project Approach

Since this is a Aalborg University (AAU) thesis, the AAU approach of problem based learning is employed. This model guarantees a problem based and structured approach to analysing a given topic, be it working with a company, or like in this case, a research field [Aalborg University PBL Academy, 2017]. The overall structure of this thesis can be seen as a funnel model, which is illustrated in figure 3.1. Using this approach, the research topic is analysed broadly in the beginning and as the thesis progresses, the analysis is getting more and more focused, just like a funnel. In the latter parts of the thesis, the funnel becomes larger again, which represents the putting the findings in a broader context again in the discussion and conclusion chapters.

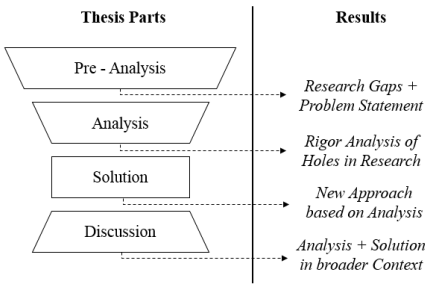

Figure 3.1: Illustration of the funnel model

3.2 Applied Methods in the individual Chapters

The overall methods that are used to create the individual parts of the thesis are described below including the results that stem from them. Although the overall methods are presented in this chapter, various other methods are introduced and applied over the course of the thesis, for example the data scraping in the solution. These specific methods are described in detail as they are introduced.

- **Pre-Analysis:** Given that this thesis is a pure research thesis, with no company involved, the pre-analysis that provides an introduction and broad overview of the field of text mining based time-series forecasting is conducted by performing a literature analysis. To do so a initiating problem statement is formulated first. Based on the

interdisciplinary of the field this analysis spans over three main fields: text mining, behavioural economics and machine learning. The result of this part of the thesis are areas, which the did not receive a lot of literature attention yet. Based on them a problem statement is formulated, that poses as the outline of the successive chapters of this thesis.

- **Analysis:** The analysis part of the thesis in chapter 6 dives deeper into the problems found in the pre-analysis and is therefore based on the problem statement. To do so, aspects that have not been considered so far have been investigated. This is, based on the theoretical nature of this thesis, naturally also based on literature review.
- **Solution:** Consequentially, a solution is developed that is predicated on the analysis part. In this section of the thesis a new framework is presented that is based on a flowchart approach [Ruth Sara Aguilar-Savén, 2004]. Subsequentially, this new approach is applied in an experiment. This experiment is conducted based on the universal workflow of machine learning, presented in the work of Francois Chollet [2018].
- **Discussion:** In the last part of this thesis, the discussion, the results of the analysis and the solution are elaborated and put into a wider context of the whole field of text mining based time-series forecasting. Since this is based on the rest of the thesis, no specific methodology is applied here.

The other chapters, like problem statement and conclusion, can be seen as complementary parts of the thesis that on one hand are the outputs of the previous parts and on the other hand are used to set the outline for the following chapter.

4. Pre-Analysis

The pre-analysis presents an analysis of the current knowledge body of the field "text mining for time-series forecasting". This analysis includes an overview a delimitation, an analysis of problem criteria that must be fulfilled in order to use these methods, a detailed description of the available framework and the methods within it and an analysis of the identified gaps in research.

4.1 Delimitation

This section presents a delimitation that narrows the focus of the further analysis. Since this is a theoretical thesis, the investigation of the field is naturally limited on the work that has been published so far. Based on the fact that text mining based time-series forecasting has not received a whole lot of literature attention, with a grand total of about 35 papers published on it, there are certain aspects that need to be considered in the analysis:

- **Sole focus on market prediction:** While the forecasting method that is the theme of this thesis is in its core universally applicable, if the application criteria are met, so far the sole focus of research is on market prediction and is therefore better known in literature as text mining based market prediction. This fact has multiple reasons, which are described later in this chapter. Regardless of this single focus on financial markets, this thesis attempts to present the whole topic in a broader context.
- **Behavioural economics:** Unfortunately, the first limitation leads to a second, which is that so far only two behavioural economic theories have been used, the adaptive market hypothesis and mood based market behaviour. This limits the analysis in this chapter and in the thesis as a whole to this two approaches.
- **Text mining and machine learning methods:** There are many text mining and machine learning methods available in literature that are almost universally applicable but in this thesis only methods are considered that have already been used in the field. The main reason for that is that otherwise the pre-analysis that introduces those methods alone would be able to fill multiple books.

Based on this delimitation the pre-analysis is conducted in this chapter, that provides an overview of the field and identifies holes in the research body as it is today.

4.2 Application criteria

In this chapter, the pre-requisites for the usage of these methods are investigated. To grasp these pre-requisites, the interdisciplinary character of this field needs to be understood first. In contrast to conventional data mining, which can be viewed as a special discipline of applying machine learning on big data, text mining for time-series forecasting is an interdisciplinary field consisting of three major components: machine learning, linguistics

and behavioural economics. [Arman Khadjeh Nassirtoussi, 2014]

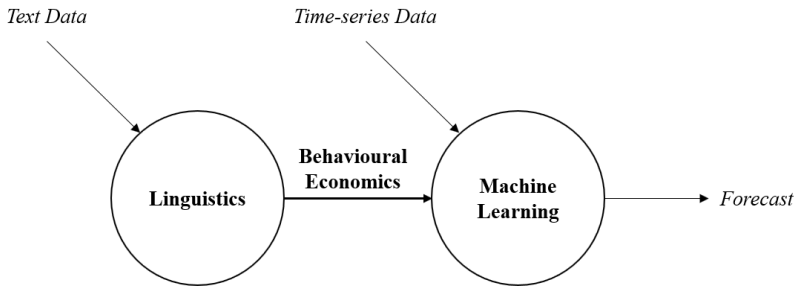

Figure 4.1: Interdisciplinary between linguistics, machine-learning and behavioural-economics

The addition of linguistics and behavioural-economics is necessary, since the methods needed to apply these methods exceed what is achievable solely with machine learning. Compared to regular machine learning methods, instead of establishing the relationship between the input data and the predicted variable mathematically, it is necessary to introduce a behavioural economic theory that establishes this connection. For the observed problem sets, these fields have the following implications:

- **Linguistics:** Methods from linguistics are needed to transform the raw text data into a format that the machine learning algorithm is able to process. That means that the unstructured data needs to be transformed into structured data.
- **Behavioural economics:** Behavioural economics establish the link between the textual data and the investigated time-series. This means there must be some explanatory theory on why the textual data influences the output variable.
- **Machine learning:** Machine learning is necessary to provide an artificial intelligence that can learn from and transform the time-series data and the (processed) text data into a forecast. This can be done by using any supervised or unsupervised learning technique like regression, neural networks, tree-based methods etc.

This interaction between the different disciplines also explains the main requirement for problems to be eligible for these type of methods. There must be a relationship between the textual data and the observed time-series, which can be explained by means of behavioural-economics. Vice versa, this also means that these methods can also be used to proof such a behavioural-economic link if they are applied to a certain problem successfully. An example for this can be found in Johan Bollen [2013], where the authors successfully establish a relationship between the overall mood on Twitter and stock prices. Further, it must be possible to convert the text data into a format that the machine learning algorithm is able to process using linguistic techniques.

4.3 System framework

To understand how text mining can be applied in the context of time-series forecasting, a general overview of the framework of the topic is needed. On a very high level, the concept

can be described as a system where:

> *At one end text is fed in as input and at the other end some market predictive values are generated as output.* [Arman Khadjeh Nassirtoussi, 2014]

In other words, this means that for forecasting problems of this kind the starting point is a standard time-series forecasting problem with the goal of forecasting the value of a given output variable. This is done by enhancing a baseline forecast done by conventional techniques by adding unstructured data as a predictive factor. The transformation of text data into a forecast is conducted in three steps, data collection, pre-processing of the data and machine learning, which consist of multiple sub-processes that are presented in 4.2. Data collection, as the name suggest, is the task of gathering and filtering the input data. In pre-processing the this unstructured input data is transformed into structured data that can be fed into a machine learning algorithm. Finally, machine learning consists of model training, evaluation and finally, the actual forecasting.

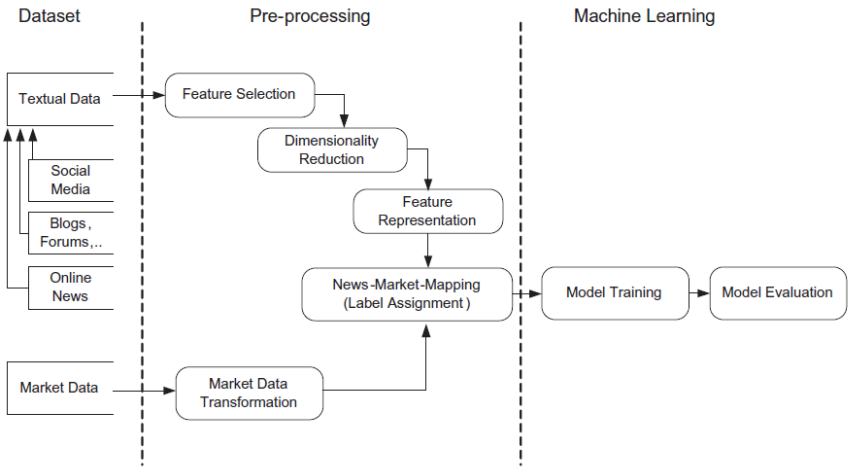

Figure 4.2: Framework of a text-mining based forecasting system [Arman Khadjeh Nassirtoussi, 2014]

In this section the purpose of all sub-processes is presented and the most common methods for each step are introduced.

4.3.1 Data collection

As mentioned above, data collection is the step in which the input data is selected and filtered. This input data consists of two data sets, the historical data of the output variable and the text data that is used to enhance the forecast. The first data set, the historical data, is given by definition and therefore is not subject to this very first step. In contrast to classical text-mining problems like classification problems, where the unstructured data is also provided by the problem definition [Sholom M. Weiss, 2010], in this case the text data is the result of some selection process. In the section application criteria, it was established that one of the application criteria of for these kinds of forecasting systems is a connection between the textual data and the output variable given by behavioural-economic reasoning. This means that when selecting the input data, it is imperative that the chosen data must correspond to these criteria. In existing literature, that focuses almost exclusively on market

prediction, the relationship between textual data and market data is mainly based on two different theories, the adaptive market hypothesis which implies that stock markets are not efficient and therefore prediction of market behaviour is possible [Andrew Urquhart, 2010] and the theory that investors' behaviour can be shaped by their overall mood [Johan Bollen, 2013]. Since there has been little to no research on other fields than financial markets, the behavioural-economic backgrounds for all other fields are yet to be established. Based on this connection, a data source can be chosen. The sources used in literature can be separated in three overall categories: social media, online news and texts from other sources like official company statements.

- **Social media:** The most common source of social media used in existing literature is the micro blogging service Twitter, where users can publish so called tweets, which are small texts with a maximum of 280 symbols (formerly 140). In Johan Bollen [2013], around 9.5 million tweets are used for their model and Vu Tien-Thanh [2012] uses around 5 million tweets for their work. Other authors use posts on message boards for sentiment analysis [Sanjiv R. Das, 2007].
- **Online News:** Another popular source of text data is online news. There are two different approaches for employing these kinds of sources, either using the whole article or just using the headlines. The main reason for some authorsto use headlines exclusively is that this leads to less fuzziness, since the headline is usually more straight to the point than the article itself [Arman Khadjeh Nassirtoussi, 2014].
- **Other sources:** Other sources that are used in literature are official statements published by publicly traded companies themselves. An example for this was presented by Michael Hagenau [2012] in his work about the influence of different pre-processing methods on prediction quality.

In addition to the selection of the input data, it may also be necessary to filter it. Again, this needs to be done in correspondence to the behavioural-economic reasoning. In the two different approaches in literature it can be observed that most presented approaches that employ the adaptive markets hypothesis as basis for their work use either online news or corporate statements and then filter it for financial news, while mood-based approaches use social media data and filter it for the published statements indicate the author's mode to some degree.

4.3.2 Pre-processing

Once the text data is selected and filtered, it needs to be transformed from its unstructured form into a structured form that can be used as input of the machine learning model. This is done in the so called pre-processing, which is conducted in three steps: feature selection, dimensionality reduction and feature representation. In classical text mining problems the choices made for the design of this pre-processing have a significant impact on performance [Alper Kursat Uysal, 2014] and the work of Michael Hagenau [2012] shows that at least the feature selection also influences the output.

4.3.2.1 Feature selection

Feature selection is the process of choosing a set of attributes that represent a given text. There is a range of different methods in text-mining literature that vary greatly in complexity and popularity. There are two overall families of methods that differ in the way they obtain features from the unstructured data.
The first group of methods obtains the attributes from the body of the text itself. Arguably

the easiest and also the most widespread technique found in literature is the so-called "bag-of-words" or single discrete words approach, which is used in around 75% of all available publications [Arman Khadjeh Nassirtoussi, 2014]. This technique simply splits a text up in its words and views every word as an attribute of the text with no regards of position within the text or special word combinations. An alternative to this very simplistic approach was employed by Michael Hagenau [2012], who implements an N-gram approach that, instead of seeing the individual words in a text as a feature, introduces sequences of words as a feature of the text. Here, the N in N-grams stands for how many words each feature contains, so the text is split up in all possible groups of N consecutive words. A different approach, which was also implemented by Michael Hagenau [2012], are so-called "Noun phrases". The basic idea of this approach is that the text's features are phrases whose head is a noun, which is optimally accompanied by adjectives.

Method	Features
Bag-of-words	The – Stock – Market – Is – Having – its – Worst – Second – Quarter – Since – the – Great – Depression
N-grams (3)	The Stock Market – Stock Market Is – Market Is Having – Is Having its – Having its Worst – Its Worst Second…
Noun phrases	Stock Market – Worst Second Quarter – Great Depression

Figure 4.3: Example of different features obtained by using different methods

In 4.3, an example for each technique is presented. As the basis for the features, the first sentence in the tweet is used. While the number of attributes obtained by using bag-of-words or N-grams is related to the total number of words of the text, noun phrases already reduce the features of the text, which is relevant for the second step in pre-processing, the dimensionality reduction.
The second family method corresponds closely to classical text-mining and uses a classification problem approach. One example for this can be found in the work of Fang Jin [2013], in-which they use "Latent Dirichlet Allocation" to classify news articles into pre-defined topics, where the topic distribution of each document is the obtained feature. A similar approach of classifying unstructured input can be found in the paper of Johan Bollen [2013], where tweets are categorised into mood states using Google's GPOMS system. GPOMS stands for "Google Profile of Mood States" and is an algorithm that is used for sentiment analysis. Another method that is a member of this family is the so-called "Named Entity Recognition" system. This method uses pre-defined entity classes, for example organisations or persons, to identify and map the boundary and the class of these entities in a given text [Vu Tien-Thanh, 2012].
The examples for the two method families above also show the main difference between the two, for approaches like bag-of-words, N-grams and noun phrases, no previous establishment of features is needed while it is imperative that there are pre-defined classes or entities when employing Latent Dirichlet Allocation, GPOMS or Named Entity Recognition systems. On one hand this makes it easier to use a method of the first family. On the other hand

this leads to a significantly larger amount of features that are obtained, since the amount of attributes is directly correlated with the number of words in the input text. This makes it necessary to reduce the number of features, which is done in the next step of the system, the dimensionality reduction.

4.3.2.2 Dimensionality reduction:

Dimensionality reduction is a crucial step in pre-processing, which aims to filter the obtained attributes from feature selection to only keep the ones that are relevant predictive factors. The main reason this process is necessary, is that it is well known that an increase in numbers of features lead to a loss in efficiency in classification and other machine learning problems, which is generally known as the curse of dimensonality [Vladimir Pestov, 2013]. This step in the system can further be divided in two sub-steps, the standardisation of the features and the filtering process itself.
Standardisation of features needs to be done first and has the goal limiting the number of features by simply reducing the number of observed words in the unstructured input data using natural language processing activities. Namely these activities are feature stemming, conversion to lower case letters, punctuation removal, removal of numbers, removal of web addresses and removal of stop words [Arman Khadjeh Nassirtoussi, 2014].

- **Feature stemming:** Stemming aims to convert the words that are part of a feature into it's root form. Instead of seeing *"running"*, *"ran"* and *"run"* as three different feature instances, they are all converted into it's root, which is *"run"* in this case by using some language dependent stemming algorithm.
- **Conversion to lower case letters:** Conversion to lower case letters is necessary since in terms of predictive power it does not matter if a word is written in upper case or lower case letters. So, if the only difference between two feature instances is if they are written in upper case or lower case, they are combined into one feature instance.
- **Punctuation removal:** Punctuation removal is necessary since words that are placed immediately before the end of a sentence or before a comma will also include those non-alphanumeric characters. So, instead of viewing *"punctuation."*, *"punctuation,"* and *"punctuation"* as three different feature instances, they are combined into one.
- **Removal of numbers:** Since numbers cannot be interpreted without putting them in context and this context is not given when using methods like "bag-of-words", in literature numbers are removed from the features in most cases.
- **Removal of web addresses:** Web addresses are removed for the same reason as numbers, as they cannot be interpreted without context.
- **Removal of stop words:** Stop words are common words that in general do not have a relationship with a certain topic and are therefore deemed to be irrelevant features of a text document in the context of text mining. Usually stop words are conjunctions, prepositions, articles etc. [Alper Kursat Uysal, 2014]

These steps are a easy intuitive way to reduce the number of features by removing features with little to no predictive value (numbers, stop words and web addresses) and combining similar features that only differ in a small way but are basically the same attribute. In this way the essential information of the unstructured data is preserved while the number of dimensions is reduced.
The second part of dimensionality reduction seeks to reduce the feature instances by only keeping the ones that are relevant to the problem itself or by keeping the ones that have

the highest occurrence rates. The methods that are found for the latter case in literature are generally fairly simple. For instance, Robert P. Schumaker [2009] and Matthew Butler [2009],introduce a minimum occurrence limit for feature instances.
A different approach is to use the features obtained in the previous step to put the text documents in pre-defined categories or classes. The number of dimensions is then limited to the number of categories introduced. This method resembles some of the techniques presented above in the section about feature selection a lot. To do so requires the introduction of an additional step, in which a connection between the text based features and the categories need to be established. Common measures to establish this connection are information gain, Chi-squared statistics and accuracy balanced, which are described below.

- **Information gain:** Information gain measures the reduction in entropy based on the result of binary representation [Serafettin Tasci, 2013]. This is a popular metric in machine learning that is often employed as an optimisation citerion in decision tree models.
- **Chi-squared statistics:** The Chi-squared statistics is a test that measures the independence of two random variables. In the text mining context, it is used to measure if a feature indicates the likelihood that the whole document belongs to a certain class [Serafettin Tasci, 2013].
- **Accuracy balanced:** Accuracy balanced is an alteration of the Chi-squared statistics, which is two sided, meaning that for each feature of a document not only the impact of making it more likely that it belongs to a certain class is measured, but also if certain features make it less likely that a document belongs to a category [Serafettin Tasci, 2013].

The third approach is arguably the most sophisticated one. It sees the process of dimensionality reduction as an optimisation problem with the object of identifying the dimensions that have the biggest predictive power. Approaching the topic from this point of view naturally leads to the development of optimisation algorithms, two examples of which can be found in Mehdi Hosseinzadeh Aghdam [2009] and Chih-Fong Tsai [2013]. Mehdi Hosseinzadeh Aghdam [2009] employs an ant colony meta heuristic in order to find a subset of the feature instances that minimises the number of feature instances while keeping the predictive value, measured either by information gain or chi-squared, at an acceptable level. Genetic algorithms as presented in the work of Chih-Fong Tsai [2013] are popular methods in classical text-mining problems and can also be employed to categorise the input documents, where the obtained class of the document is seen as the obtained feature instance.

4.3.2.3 Feature representation:

The third part of the pre-processing of the unstructured input data is the so called "feature representation". After the features have been obtained in feature selection and the irrelevant features have been filtered out in dimensionality reduction, this step converts the remaining information into structured data that can be used by the machine learning algorithm. That means that for each document a numeric value has to be obtained for every possible feature. The most common techniques to do so are binary representation and term-frequency-inverse document frequency [Arman Khadjeh Nassirtoussi, 2014].

- **Binary representation:** Binary representation is the simplest technique, which also makes it very popular and widely employed in literature. The basic approach is that for each document a binary value, either 1 or 0, is assigned to every possible

feature, depending on the existence of the feature in the observed document.

- **Term-frequency-inverse document frequency:** The term-frequency-inverse document frequency measures the frequency of the occurrence of a feature within a document compared to the frequency of the feature in the whole input data [Arman Khadjeh Nassirtoussi, 2014].

With this step being completed, the pre-processing of the text data is finished, and the result of the process can be fed into a machine learning algorithm.

4.3.2.4 Simple example of text pre-processing:

The whole pre-processing requires a lot of decisions regarding the use of methods and they can vary a lot in how they transform the data, which makes it hard to imagine how the data is processed throughout the process. To provide a clearer view of how the data is transformed, this sub section shows an example for such a process, using some of the simplest of techniques for each step, which are "bag-of-words" for feature selection, "minimum occurrence limit" for dimensionality reduction and "binary representation" for feature representation.

Input Data	
Text #1	The Stock Market Is Having its Worst Second Quarter Since the Great Depression
Text #2	FYI, the Stock Market hasn't had it's second worst quarter since the Great depression because of something Obama did.
Text #3	If Trump is going to take credit for stock market gains, he also must own the worst April the stock market has seen since the Great Depression.

Figure 4.4: Text Inputs for pre-processing example

Figure 4.6 shows the texts that are used as the input for this example. These texts are three tweets that were posted about the same overall topic and were published exactly as they are presented above. Using the bag-of-words method the features of the three texts are simply the individual words within them.

Dimensionality Reduction	
Text #1	stock market worst second quarter since great depression
Text #2	stock market second worst quarter since great depression something obama
Text #3	trump going take credit stock market gains must worst april stock market since great depression

Figure 4.5: Dimensionality reduction of the input data

As it was established above, the number of features must be reduced in the dimensionality reduction step. First, the three texts must be standardised by convertion to lowercase letters, removal of punctuation and stop words and word stemming. The result of this process can be seen in figure 4.5. Only with this standardisation was the number of features of text 3 reduced from 27 to 15. Next, a minimum occurrence limit of each feature is introduced, which is set to three in this example. That means that a feature must occur at least three times in the total input corpus to be considered as a feature and all other words are ignored. This leads to the following list of attributes: *"stock"-"market"-"worst"-"second"-"quarter"-"since"-"great"-"depression"*.

Feature Representation								
Features	stock	market	worst	second	quarter	since	great	depression
Text #1	1	1	1	1	1	1	1	1
Text #2	1	1	1	1	1	1	1	1
Text #3	1	1	1	0	0	1	1	1

Figure 4.6: Feature representation of the input data

In a final step, the features of the texts are converted into a numerical representation using binary representation, meaning that for each feature either a 1 or a 0 is assigned to the input text depending on the existence or absence of the word in the text. Using these example techniques, the result of the pre-processing is a vector that has the same length of the number of features selected in dimensionality reduction, in this example 8. For instance text 3, *"If Trump is going to take credit for stock market gains, he also must own the worst April the stock market has seen since the Great Depression."*, is now simply represented by the vector (1 1 1 0 0 1 1 1), which can be used as input data for the next step in the system, the machine learning algorithm, after it has been mapped to the historical data of the prediction variable.

4.3.2.5 Text - time series mapping:

The final part of the pre-processing process is the mapping of the now structured textual input to the time-series data. The purpose of this step is to establish a time-based relationship between the two types of input data. This is done in two steps: first a time line is introduced in which the values of the predicted variable and the input documents are mapped based on their release data. In a second step, a time interval has to be introduced that defines which text inputs must be grouped together. Accordingly, for every point *"x"* on the time line a certain interval *"Y"* is defined and all documents which have a release date in *"Y"* are mapped to *"x"*. It is important to note that in most of the cases the end time of *"Y"* will not be *"x"*, but there will be some lag in between the two [Arman Khadjeh Nassirtoussi, 2015]. Therefore, two parameters have to be introduced, the lag and the width of the interval. The reason for that is based on the underlying behavioural-economic component of the observed problem, which in most cases indicates that news, tweets etc. will not have an immediate impact on the predicted variable, but it takes a certain amount of time to process unstructured information, after which its influence will start to affect the observed time series.

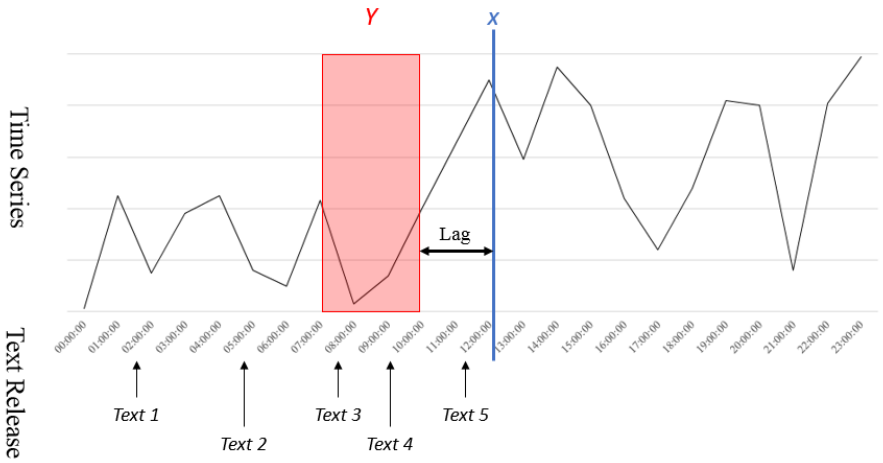

Figure 4.7: Example for the mapping of the textual input to a time series

In figure 4.7, an example of this text to time series mapping is shown. Here *"x"* represents the observed point of time and *"Y"* is the interval in which texts are mapped to to the observed predicted variable at point *"x"*. In the shown instance, only texts 3 and 4 are mapped to *"x"*. Text 5 is not mapped to *"x"*, even though it's publication date was before *"x"* since not enough time has passed to process the information in the input. This process needs to be done for every observation in the time-series and the finished mapping is the final input of the next step, the machine learning algorithm.

4.3.3 Machine learning:

The machine learning algorithm is the mechanism that is used to combine the processed information obtained from the unstructured input data and the historical data of the prediction variable. In the purest form of text-mining based time series forecasting the only structured input is the historical data of the predictive variable, employing an approach based on auto correlation, but additionally, other structured input can be used like in every forecasting system. In literature, different methods have been used, but by far the most prominent algorithm is the "support vector machine" (SVM) approach, which has been employed in about half of the cases [Arman Khadjeh Nassirtoussi, 2014]. Other methods that can be found in literature are regression models [Fang Jin, 2013], decision trees [Vu Tien-Thanh, 2012] and neural networks [Johan Bollen, 2013]. In this section the usage of these popular methods in the context of text-mining based time-series forecasting is described including the reason of their usage and their possible advantages and disadvantages.

4.3.3.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning classifier algorithm that can be applied for two class settings, meaning that the output of the method is binary. In the context of text mining based time-series forecasting, this means using a SVM not a actual outcome of a time series is predicted, but rather if the next observation is going to be higher or lower than the current one, with no regards of how significant the difference is. The underlying principle of this technique is that the predicted variable is part of a feature space, and, depending on the region of the feature space the observation is located in, an estimation can be made about the likelihood of the observation belonging to either of

the two classes. To do so the boundaries of the region in the feature space have to be identified. In more basic methods like "support vector classifiers", these boundaries are linear by definition, while SVM finds non-linear boundaries using so called kernels. The calculation of these kernels is quite mathematical and requires some knowledge of linear algebra. The feature space that was introduced above can be viewed as an n-dimensional space, where n equals the number of existing features. In this space, every observation can be represented by an n-dimensional vector, where the elements of the vector are simply the observations features.
Like mentioned above, to find the boundaries in these spaces, kernels are introduced, which are functions that quantifies the similarity between two observations. The simplest kernel is the *linear kernel*, which basically uses the inner product of two observations, which is defied as

$$K(x_i, y_i) = \sum_{j=1}^{p} x_{ij} y_{ij} \tag{4.1}$$

where x_i and y_i are the two observations and j is the number of features. This results in a the class-prediction function

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{4.2}$$

where α_i and β_0 are factors that are decision variables that have to be determined using some training data set S, where $x_i \in S$.
To escape the problems that arise with the usage of a linear boundary, other kernels can be used, for example the *polynomial kernel* and the *radial kernel*, which lead to a more flexible decision boundary. The formula for these two kernels are

$$K(x_i, y_i) = (1 + \sum_{j=1}^{p} x_{ij} y_{ij})^d \tag{4.3}$$

where d is a positive integer, for the polynomial kernel and

$$K(x_i, y_i) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - y_{ij})^2) \tag{4.4}$$

where γ is a positive constant, for the radial kernel Gareth James [2013].

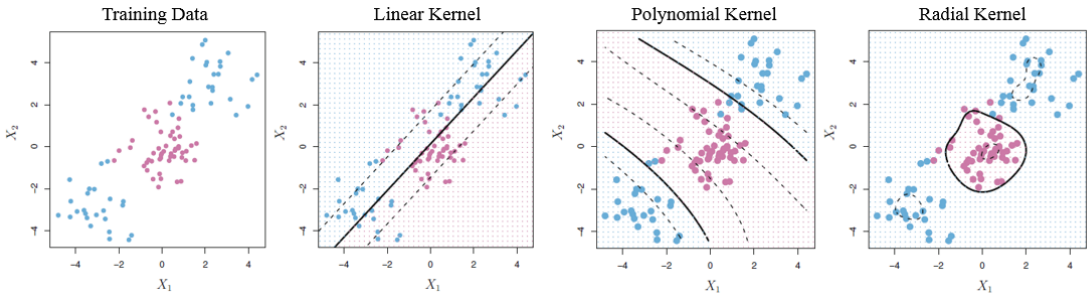

Figure 4.8: Example for different kernels on the same training data (left) from Gareth James [2013]

In figure 4.8, the three different kernels are applied to find the boundaries for the same training data set that is shown in the very left of the figure. In the training data set the

two different categories of observation are visualised by the colours of the observed points, either blue or violet. Since the violet observations are located in between two clusters of blue observations, the linear kernel is performing rather poorly in finding a boundary in between them, while the two kernels associated with SVM, polynomial and radial, are performing quite well.
One of the main reasons that SVM is such a popular model in text-mining based time series forecasting is that the computational effort is not increasing significantly with a larger number of features but rather on the size of the training data Arman Khadjeh Nassirtoussi [2015]. This is especially important in this context since this provides room for a larger set of features obtained by the method chosen in dimensionality reduction. Further, it is known to perform rather well, which possibly is based on the binary outcome of the method. Although this might lead to better performance, it also limits the problems it can be applied to. While the vast majority of literature in this field was done on financial markets, where it is sufficient to predict an up- or down movement, with no indication of how large this movement is, using this method, there are other fields where this might not be enough. In such cases, SVM cannot be used.

4.3.3.2 Regression models

To avoid this conflict, some authors like Fang Jin [2013] or Robert P. Schumaker [2009] employ regression models. There are two main classes of regressions that are used, classical regressions like multivariate and linear regression, and an alteration of the SVM method that was introduced earlier, the so-called "support vector regression" (SVR).
In the instances in which classical regression is used like in Fang Jin [2013] and Paul C. Tetlock [2008], the basic idea is just to use the product of the pre-processing model in an regular regression model. For example, Fang Jin [2013] uses the methods in the pre-processing to categorise news articles into their topics and then maps them using the changes in coverage of the individual topics from time interval to time interval as the inputs for the regression function, which uses these inputs to calculate the currency fluctuation. The authors method can be represented with the simple linear function

$$\Delta c_{t+1} = \beta_r \Delta r_t + \beta_f \Delta f_t + \beta_e \Delta e_t + \beta_s \Delta log(s_t) \tag{4.5}$$

, where c is the currency, r is the change in articles about interest rate, f is the change in articles about inflation rate, s is the change in articles about stock markets and e is a dummy variable that represents unexpected events. The model is then optimised on the previous two weeks to obtain the β.
A more advanced regression that has been used in literature is the SVR, which was used by Robert P. Schumaker [2009]. The basic idea follows the principle of SVM closely, where to goal is to use training data to find the boundary between the two classes. In contrast to SVM, SVR is not a classification method but rather seeks to find the function in the vector space that does not minimise the sum of errors for all observations (like the least-squares regression) but keeps the error of each individual observation below a pre-defined value. Since this is not feasible for most problems if the output function is required to be linear, kernels need to be introduced here as well. Once this function is found, the distance between a new observation and the test observations (so-called "support vectors") can be calculated, and therefore, the position on the function and the value prediction can be obtained [Alex J. Smola, 2004].

4.3.3.3 Decision Trees:

Another machine learning method that is used in literature are so called "decision trees", which are for example employed by Vu Tien-Thanh [2012] in their work. The idea behind decision trees is to segment the solution space into small areas and then assigning a forecast value or category to an observation by simply evaluating in which of the areas it is located. To convert this assignment to a certain area in the solution space into a prediction simply the mode or the mean of the historical observations in the area the observation belongs to is used.

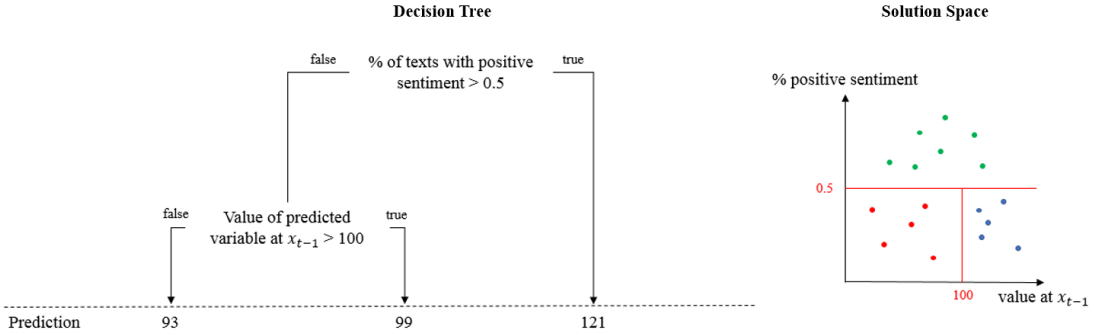

Figure 4.9: Example for a simple decision tree

Figure 4.10 shows a simple example of how a decision tree works. The example is based on a text-mining based forecasting system in which the input texts are categorised into positive and negative based on their sentiment. The decision tree consists of two branches, first the percentage of positive texts in the feasible interval described in the text time series mapping section, for the observation and the previous value of the predicted variable. If a prediction is to be made for a new observation, first it is checked if the percentage of texts with a positive sentiment is larger than 0.5. In this case the prediction would be 121, which is the mean of the green observations in the solution space. Otherwise the previous observation is checked and if that is larger than 100, then the mean of the blue observations is used. Else, the mean of the red observations is used as a prediction. This example leads to the question, how to split the solution space. In general, this is done by finding the split that minimises the total residual sum of squares (RSS) for all observations. It is not feasible to find a combination of splits that does that in one step, since if continuous features are used the number of possible combinations would be infinite. This is done in a top-down, greedy approach where one split is found at a time [Gareth James, 2013].
Using this method for text-mining based time-series forecasting is not particularly popular but it has one advantage, it can be used both to predict categorical and continuous outcomes. That means that it can be used to only forecast an increase or decrease of the predicted variable or to provide a numerical forecast.

4.3.3.4 Neural Networks:

Another machine-learning method that can be found in literature, for example in the work of Johan Bollen [2013], are neural networks. Neural networks have gotten a lot of research and public attention with the surge of artificial intelligence in recent years, where these methods belong to the sub-class of deep learning. The deep in deep learning stems from the fact that those methods process the provided input data in successive layers of

representation. The purpose of these layers is that each one transforms the input data into representations that differ more and more from the original input with each layer and every deeper layer is supposed to provide more information about the final result. In simple terms, a neural network seeks to simply transform the data that is fed into the model. A neural network consists of multiple elements: the above-mentioned layers, certain weights that are assigned to the layers, a loss function and an optimiser to find the weights based on a training dataset.

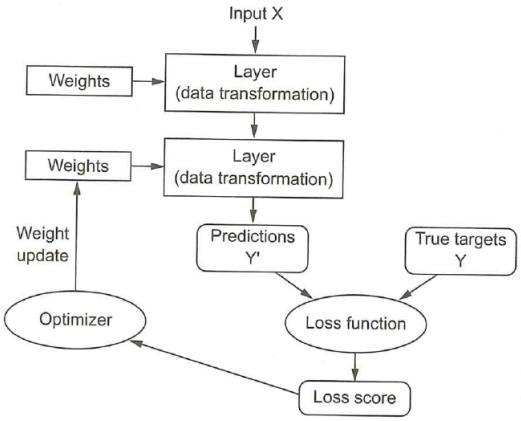

Figure 4.10: Anatomy of a neural network [Francois Chollet, 2018]

The first element are the layers which consist of so-called neurons. In every layer the neurons can either get activated or not based on the information that the layer gets from the previous layer (or the input data itself in case of the first layer) and the assigned weights. These weights indicate the relationship between layers, i.e. how the output of the previous layer influences the next layer's neurons activation. These weights are also called parameters and are the decision variables of the whole system that are adjusted in order to optimise the performance of the neural network. This optimisation is done within a pre-defined training dataset with the introduction of a loss-function and an optimiser. In this training data, the desired output is known, and the neural network's predictions are compared to these true targets. The distance between the predictions and the desired values is then calculated by the so-called loss function. In a next step, the weights of the network get adjusted so that the neural network performs better and better, which is normally done by the optimiser using so called backpropagation. When the loss function is minimised, the neural network provides results for the training set, which are as close as possible to the true targets. For this reason, it is called a "trained network". [Francois Chollet, 2018]

4.4 Problem areas

The sections above introduce the field of text mining based time-series forecasting including the application criteria and the system framework with the major methods that are used in the field. This in-depth introduction covers the vast majority of the literature that has been published so far, although it has to be said that the research body is not very big since it is such a young field that has only gotten some research attention recently. In this section about the problem areas the holes in the literature are identified and it is shown where the problems within the research body lie.

4.4.1 One sided research focus

The most obvious problem with the research that has been done in this field is one that can be found in many research disciplines, which is that the work that has been done so far has had a very narrow focus on the application of these methods while very little work has been done on gaining a deeper understanding. Consequently, the vast majority of the popular works in research (around thirty papers) has its main focus on identifying a problem that is feasible for these methods and introducing a solution approach for that specific problem, while there are only three relevant papers that focus on a more holistic view of the field, Arman Khadjeh Nassirtoussi [2014], Azadeh Nikfarjam [2010] and Michael Hagenau [2012]. In their paper that was published in 2014 Arman Khadjeh Nassirtoussi [2014] reckon:

> *Despite the existence of multiple systems in this area of research we have not found any dedicated and comprehensive comparative analysis and review of the available systems.*

Although this simple focus on applicable cases can be found in multiple research areas, in this field it is especially severe at this time since the field is still young and there simply hasn't been a lot of research done yet, which consequentially leads to this lack of a "holistic" view on the field. This fact combined with the interdisciplinarity of the area, which makes fundamental research even harder, creates an incentive to focus on applications. Furthermore, it simply takes more effort with less chance of success to do basic research in this field, which makes appliance-based research even more attractive. In their recent paper on the state of machine learning D. Sculley [2018] state:

> *While the pace of progress has been extraordinary by any measure, in this paper we explore potential issues that we believe to be arising as a result. In particular, we observe that the rate of empirical advancement may not have been matched by consistent increase in the level of empirical rigor across the field as a whole.*

Although this one-sided literature focus is understandable from a researching point of view, it consequentially leads to a lot of missing knowledge. Maybe the most obvious effect is that so far, the whole research body deals exclusively with market prediction. There is not a single paper that focuses on anything but market prediction, although the system framework would easily allow that. There are multiple possible reasons for that, one being the one that was mentioned above that it is simply attractive for researchers to conduct work with a high chance of success. Another possible reason lies in the interdisciplinary nature of this area, which requires a behavioural economic connection between the unstructured data and the predicted output variable. This is given in market prediction with the adaptive market theory, which has received a lot of research focus over the years and is well established. This leads to the next problem area in the field, the lack of understanding of the behavioural economic component.

4.4.2 Overall complexity of the framework

The fact that this field is so young and there is not a real extensive research body on it, especially not basic research, amplifies a problem that can be found across the whole field of machine learning, namely that while a lot of methods can be viewed as a "black box", the whole field becomes more and more of a "black box". In a recent talk, one of Google's

AI researchers even went so far to compare machine learning to alchemy, where it is known in general that it works because the results prove it, but it is not understood why and how it works [Hutson, 2018].
While this is already a problem in regular machine learning where a system typically only consists of two steps, data collection and machine learning, this is especially true in text-mining based time series forecasting since the whole system framework is so much more complicated. Instead of simply going through two steps, the text must be processed and the different methods that have been presented above need to be implemented and applied. This leads to a whole new "complexity dimension", which is not understood, that the effect of the combination of the text processing methods is in no way known. Basically, instead of having one "black box", the machine learning part of the system, a second "black box" is added, which is the text processing part. Additionally, there is a need for pre-defining certain parameters in the system, where the effect of the decisions made in that part of the system are also unclear. This leads to a large uncertainty across the whole decision making in the design process of the system. In other words, if the whole forecasting system is seen as a machine where the methods with their sub-processes are seen as the gears of the machine, it is possible to observe a change in the end result that the machine produces when we change one gear with another, but neither can the change in the result be anticipated nor can the influence of the gear change on the other gears be observed.

4.4.3 Behavioural economics

As was described in the application criteria section above, there are three major components that make up the framework of text mining based time-series forecasting, linguistics, machine learning and behavioural economics. While the first two are straightforward in that they provide the methods that are used in the system framework presented earlier in this chapter, behavioural economics is fuzzier in that it has to provide the logical connection between the text data and the output variable. As mentioned above, this connection is well established in market prediction with the adaptive market theory. Anyhow, in other appliance fields it might be more difficult to find such a connection, leading to the need of establishing such a connection beforehand, which can prove to be rather difficult.
When following this approach, the whole forecasting system is based on an initiating hypothesis about the relationship between the unstructured data and the output variable. The dilemma in this case is that there is no way to prove such a relationship beforehand like it would easily be possible when working with structured data. That means that the only way to prove the hypothesis is to construct a forecasting system that provides a prediction quality that is significantly improved over a "pure" machine learning solution without the supporting text mining. And even if this proves to be the case, the details of that relationship still remain fuzzy. A example for this approach can be found in the work of Johan Bollen [2013], in which the hypothesis is made in the beginning that the overall mood in the United States is linked to the Dow Jones Index, based on the theory that financial decisions are significantly driven by emotion and mood [John R. Nofsinger, 2005]. This hypothesis is then tested and to be "proven" by the development of a text mining time-series forecasting system that obtains the overall sentiment from tweets and then links them to movements in the Dow Jones. And even though this theory might not have been established before, it still had a solid foundation and the part of the hypothesis that was not clear beforehand was if the mood that can be obtained from twitter is correlated to the overall public mood or at least the sentiment of market participants trading the Dow. In conclusion the development of a hypothesis like this that connects text to time series data can be a daunting task, which is one of the main reasons why this method is so

far only applied for market prediction.
Another problem that is linked to the behavioural economic part of the system is that it actually has an influence on the design of the forecasting system, meaning that for each underlying hypothesis there are certain methods that become infeasible if one or another theory is chosen as the connection between text and time series data. Even though there are only two basic theories available in literature so far, selecting either one of them rules out some approaches in system design. The mood-based approach is such an instance, where in the feature-selection part of the system it becomes invalid to obtain the features from the text body itself, but the textual input has to be classified by definition, since mood is a categorical variable by default. This fact has found no attention in research so far, and therefore, very little knowledge about the influence of the behavioural economic base on system design exists.

4.4.4 Relationship between the layers of the forecasting system

While the missing knowledge about the connection of the behavioural economic logic of the system and the system design is a problem, it is also rather problematic that very little is known about the influence of choices in one layer of the system on the design of the following layers. This problem is also based on the prime issue of this research field that was introduced earlier, the one-sided research focus. In the publications so far, it is almost exclusively the case that the system was designed step by step where in every layer exactly one method was chosen that was working in the case, with very little regard of how that method interacts with the layer above or below or even reasoning why this method was chosen. As a consequence, very little is known about first of all what the influence of the decision on the overall system and what it means for the decision making in the following layer. What is meant here is that it is already problematic that there is little to no knowledge about which combination of methods works best, it is not even known which methods can be combined at all and how they can be combined.
In the subsection about the system framework, the overall system design with all its layers has been introduced and an extensive overview of the different approaches in every layer including a description of the methods is provided. This is based on what has been done in literature so far, where like mentioned above, most works in research use one set of methods. While this means that it is known which specific combination of methods in the different layers are feasible, there is no work so far on which combinations are possible in general and maybe even more important, which methods are not working. This fact is rather problematic for anyone that designs a text mining based time-series forecasting model, since it makes it hard to make an educated decision on which methods to choose in every step. If there is no such design framework available, the danger exists that in every step simply the method is chosen that fulfils the previous layer's requirements rather than the method that makes the most sense.

4.4.5 Impact of the nature of the forecasting problem itself

While the normal process in the real-world is that a very specific well-defined problem needs to be solved, this is not the case in research where problems are posed as either finding a solution to a generic class of well-known problems (e.g. the job-shop scheduling or the travelling salesman problem) or filling a void in research by attaining a new insight. In the context of text mining based time-series prediction this means in a company a certain problem would be present, for example "How can the daily exposure to currency exchange rates be minimised?", which needs to be solved. Given this problem, the first

step is to ensure that an approach using the method would ensure the application criteria, which is the case for the given example since foreign-exchange-markets are subject to the adaptive market hypothesis [Christopher J. Neely and Ulrich., 2009]. Therefore a behavioural economic basis is given, and a forecasting system which fits the criteria of the problem can be developed accordingly. Unfortunately, it seems like the literature does it the other way around. Instead of finding a solution that fits the problem, a problem is found that fits the solution and more importantly, the available data. This approach can be found in the majority of the papers, for example in the work of Matthew Butler [2009], who tests how company statements can be used to forecast the companies' stock prices or Vu Tien-Thanh [2012], who simply forecasts the stock prices of four publicly traded enterprises.

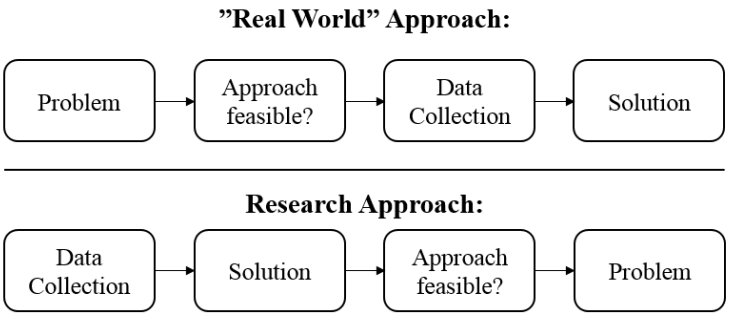

Figure 4.11: Different approaches in research and industry

The inverse problem solving approach in research seems illogical at first glance, but there are some valid reasons for it. First and foremost, it cannot be understated how hard (and possibly expensive) it is to retrieve relevant data. Although the sheer amount of data is almost rising exponentially, in today's world data is a valid currency and the enterprises that generate and collect it are very protective of it. Second, the fact that the research body deals exclusively with market prediction leads to the fact that finding a problem is not as important as in other field since financial markets have an intrinsic problem, maximising profit while limiting risk.
While the reasons of researchers are understandable, this approach nevertheless leads to the issue that the impact of the specific problem that needs to be solved remains unknown.

4.4.6 Summary

This chapter, the pre-analysis, provides an overview about the whole research field of text mining based time-series forecasting. First, the literature that has been done in the field and the history of the field was presented briefly, then the application criteria was introduced including the connection of the interdisciplinarity of the field with linguistics, machine learning and behavioural economics, which need to work together. Next, a overview of the whole system framework was presented with the purpose of every step that needs to be conducted and a description of all the popular methods for each step. Finally, the problems with the current state of research were shown in the problem section of the chapter.
These problems can be summarised shortly in one main problem, the one-sided focus of the research, and problems that stem from that.

- **One sided research focus:** So far research has almost exclusively focused on applying this method of time-series forecasting on concrete real-world problems and

very little work has been done on base research, which is typical for machine learning.

- **Overall complexity of the framework:** This main problem is amplified by the fact that in contrast to "normal" machine learning problems, where only one "black box", the machine learning itself, is present, a second "black box" needs to be introduced which is the text processing part of the system.
- **Behavioural economics:** Additionally, the behavioural economics component of the whole field makes the whole framework more complicated since it has some influence on the system design, which has not been investigated so far and makes it harder to apply this forecasting method to a wide area of problems.
- **Problem impact:** The pre-analysis further shows that a specific problem that is supposed to be solved by text mining based time-series forecasting possibly also has influence on the system, which has not been a topic in research so far.
- **Relationship between the layers of the forecasting system:** Because of the one-tracked style of research so far, there is very little knowledge about how the chosen method in one layer of the system influences the possibility of choosing methods in the following layers. Moreover the influence of a chosen method on the final result and other methods is not clear, which is called the "Winner's curse" in D. Sculley [2018].

Based on these identified problem areas a problem statement is formulated in the next chapter of the thesis that outlines the main topic and the rest of the thesis.

5. Problem Statement

A comprehensive overview of the the field text mining based time-series forecasting was presented in the pre-analysis in chapter 4. Based on the initial problem statement, which was to identify the gaps in literature in the field, the current research body was analysed and the main problem was identified, the overwhelming focus of research on the application of the technique. Subsequently, this leads to more problems in understanding the field, based on the lack of knowledge about the influence of the behavioural economic part on the system design and also about the relationship between the different layers of the framework. Additionally, these problems are amplified by the overall complexity of machine learning problems, where the text mining based time-series forecasting must be counted as one of the more difficult problems based on the interdisciplinarity and the sheer number of methods that need to be applied. In order to investigate the influence of the behavioural economic component on system design and the limitations in the choices for the framework design based on methods that were chosen already, further analysis will be conducted for these problems. Therefore, the problem analysis will answer the following problem statement:

Which factors influence the design of a text mining based time-series forecasting problem?

In addition to investigating this problem, another research question is formulated and investigated based on the problem of the overall complexity of the field:

- How does the complexity of the system influence the forecasting system?

The next chapter of the thesis, the problem analysis, provides a comprehensive breakdown of these questions.

6. Problem Analysis

This chapter is based on the problem statement in chapter 5. The purpose of this chapter is to analyse the problems that were found in chapter 4 in more detail. Through this analysis, it will be determined how the behavioural economic component influences the possible choices for methods when designing the system, how the selected methods in each layer influence the feasibility of the methods in the successive layers and if there is a way to reduce the complexity of the system. This chapter will begin with an investigation of the two known theories used as the behavioural economic component and if and how each of them influence the possible system design. Afterwards, the relationship between the different layers of the system framework is analysed and it is reviewed if and how they interact with the rest of the framework. Finally, the complexity of the current framework is examined and if there are alternative approaches that possibly reduce the complexity.

6.1 Influence of the behavioural economic component

As described in chapter 4, the behavioural economic component is one of the three building blocks every text mining based time-series forecasting system is built upon and it is the link between linguistics and machine learning. The basis of this analysis is the two different theories that have been used so far, the adaptive market theory and the influence of mood on markets. The foundation of these two theories is somewhat similar, where both the adaptive market theory and the theory of mood influenced markets claim that the markets are ineffective (meaning that markets are not subject to random walks, which would make them impossible to predict [Burton G. Malkiel, 2003]). In the following sections, both of these approaches are described in more detail and their effects on the design of text mining based time-series predictions are described.

6.1.1 The Adaptive Market Hypothesis

For most of the 20th century the efficient market theory was the gold-standard of the understanding of financial markets. It implied that financial markets fulfil the criteria of having perfect information and that all market participants will always act rational based on this information making the markets efficient, which means that all price movement are approximately subject to a random walk [Burton G. Malkiel, 2005]. This theory was more and more challenged from the late 20th century onward by behavioural economists, who criticise this hypothesis based on the idea that humans are simple not so called "homines-oeconomici" that always act rational but rather are influenced by their perception of reality and their emotions. This criticism is also somewhat supported by the fact that a lot of the champions of the efficient market theory left science to found hedge-funds that acted exactly opposite of what they claimed would be effective in the market by trying to forecast stock prices and engaging in trading methods like stock picking [Sebastian Mallaby, 2010]. Based on this discrepancy between the two opposing views on markets, a new theory

has been developed that is based on the efficient market theory but incorporates principles of behavioural economics, the adaptive market hypothesis. According to Andrew W. Lo [2005], the primary components of this new approach are:

- Individuals act in their own self interest
- Individuals make mistakes
- Individuals learn and adapt
- Competition drives adaption and innovation
- Natural selection shapes market ecology
- Evolution determines market dynamics

Especially the second and third point vary from the efficient market theory and means that markets are inefficient, since a system in which mistakes are made and adaption happens can by definition not be efficient.
One way to use text mining based time-series forecasting is to follow this hypothesis. First of all, its core message that markets are ineffective mean that they are not subject to random walk and therefore can be forecasted with the right methods. Second, the view that markets are subject to interaction with learning and adapting individuals that are prone to mistakes consequentially mean that the process in which the market participants process information varies from individual to individual. This is the starting point for the forecasting method that is the subject of this thesis, evaluating unstructured information and how individuals act upon it.

6.1.1.1 Adaptive Market Hypothesis and the influence on the system design

Since the adaptive market theory is rather open and explains more the overall behaviour of market participants, the adaptive market theories influence on the design of the forecasting system is rather weak. Broken down on the individual steps of the text-processing part of the forecasting system this means:

- **Data collection:** The Adaptive Market Hypothesis leaves a large degree of freedom for the data collection process. The only criterion that needs to be fulfilled is that the data is processed by individuals in the market and therefore is subject to their interpretation, part of their learning process or influences them in some other way. This legitimises all data sources that can be found in literature, no matter if it is news, blog posts or social media posts, since they all have an influence on either the market view of the market participants or their general behaviour.
- **Feature selection:** As presented in chapter 4, there are two main groups of methods for feature selection, one where the features are obtained for the text body itself and the other where the features are the classification of a whole text document. Both variants are legit in regards to the behavioural economic background since both types of information have an influence on the individual.
- **Dimensionality reduction:** Similar to feature selection, the openness of the Adaptive Market Hypothesis leaves a large degree of freedom for the dimensionality reduction. Further, since the dimensionality reduction is a rather technical step that is based more on mathematical methods, the influence of the behavioural economic foundation of the system on this step is rather low.
- **Feature representation:** Since the feature representation is more of a technical step than one that depends strongly on the Adaptive Market Hypothesis, it has no real influence on this step either.
- **Machine learning:** The Adaptive Market Hypothesis can have some influence that is related to the selection of a machine learning model, but it is not a direct

relationship, as it allows methods to be chosen that are (binary) classifiers, since it will always be used for market prediction.

Summarised that means that the Adaptive Market Hypothesis provides a large degree of freedom for the system design, and basically all methods that were presented in chapter 4 can be chosen. This is based largely on two main reasons, first that this theory is not very specific in how information influences the market participants but rather only claiming that there is a relationship, and second that a vast majority of the research done so far in the field is based on this hypothesis, meaning that by default most to all of the methods that have been employed so far must be applicable using this behavioural economic basis.

6.1.2 Mood based market behaviour

The second behavioural economic foundation that can be found in literature is the hypothesis that the market participants overall mood influences their actions and therefore market movements. It is quite obvious that this theory is not contradictory to the Adaptive Market Hypothesis, but rather tries to identify one factor that influences the characteristic individuals that are described in 6.1.1. This theory was first presented by Johan Bollen [2013] who claims:

> *Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role (...) Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news.*

So, while the Adaptive Market Hypothesis focuses more on how individuals act, Bollen et al's hypothesis focuses solely on how they are influenced by their emotions. The quote above shows that in this hypothesis, while it is acknowledged that news items are relevant, they are taken out of the equation and emotions are the only factor that is observed. Since it is close to impossible to live track the emotions of the market participants, Johan Bollen [2013] choose to track the mood of the whole United States using tweets.

6.1.2.1 Mood based market behaviour and the influence on the system design

While the Adaptive Market Hypothesis provides a very open framework, the mood based market behaviour theory is more restricted in what the predictors for market movements are in that it completely excludes news and only focuses on emotions. Using this behavioural economic basis restricts the possible choices in system design significantly. The effects on the individual steps in the text processing design are the following:

- **Data collection:** The mood based theory limits the possibilities of which data to use decisively. Since the only valid predictor is emotions it is imperative that the information about emotions can be derived from the data that is collected. This means that whatever data is collected must be feasible input data for sentiment analysis and opinion mining methods, which use sophisticated techniques to extract the feelings that are connected to text like the one presented in the work of Alexander Pak [2010]. A valid source for such data are social networks like Facebook or micro-blogging services like Twitter, which is also arguably the most popular one.
- **Feature selection:** Using this behavioural economic basis has a severe effect on the

methods that can be used for feature selection. As described in 4, there are two main groups of techniques for feature selection, one that is based on the text's body itself and one that is based on classifying the text. In this case the first one cannot be used since the count of a word, a combination of words or a noun phrase itself without a classifier does not provide any information about the emotion embedded in the text. So in this case, only classifiers can be used which assign a certain value for different emotions to a text document, for example the Google's GPOMS algorithm used in Johan Bollen [2013]. Sub-sequentially, it is also necessary to define beforehand what the classes (or the emotions) are that are subject to classification.

- **Dimensionality reduction:** The limitation of applicable methods in feature selection further influences the feasible methods in dimensionality reduction. Since the text is already classified and the number of features is pre-defined by the classes that were introduced in the feature selection, in this case it might not even be necessary to use dimensionality reduction. If it is chosen to reduce the dimensions anyways, the simpler methods like minimum occurrence limit are not feasible but the reduction has to be seen as an optimisation problem where the classes with the largest predictive power need to be identified.
- **Feature representation:** Although the influence on the other layers is rather high in the mood based approach, the feature representation is relatively uninfluenced by it since both popular methods can still be used, either the count of occurrences of one class or the relative frequency.
- **Machine learning:** Just like the feature representation, the machine learning part of the system is not influenced by choosing a mood based approach.

While the Adaptive Market Hypothesis leaves a large degree of freedom for the system design, the mood based market behaviour theory is quite limiting in some parts of the framework, especially in data collection and feature selection, where some methods or sources simply become unfeasible. This is based on the very specific nature of the hypothesis itself, that is way more narrow than the hypothesis investigated above.

6.1.3 General influence of the behavioural economic component

The two examples above show that the choice of a behavioural economic component can have severe influence on the rest of the forecasting system. While only the two concepts presented above have found attention in literature so far, there are some conclusions that can be derived from them that are universally valid.
One of the main findings in comparing the two is that the opener the chosen behavioural economic basis is, the more choices are available in each layer of the system framework. While this can be seen as an advantage in general, it also means that there is more uncertainty about the performance of the system to begin with. If a very specific basis is chosen that describes a very specific part of the relationship between the text data and the predictive variable, it becomes more likely that a system based on it is successful, given that the theory has been proven to be true before. In contrast, a very open basis like the Adaptive Market Hypothesis provides a large degree of freedom in system design, which also leads to a larger uncertainty about the efficiency of the whole system since the nature of the relationship between input data is not really known.
This finding holds true for all fields in which text mining based time-series forecasting is applied. The more specific the relationship between the text data and the predicted variable is formulated based on the behavioural economic background, the more restricted the system design is, but a proven hypothesis also means a higher certainty of success, at least in theory. Additionally, it has to be recognised that while it is the case for mood

based market behaviour, it is not necessarily the case that a very specific behavioural economic basis limits the choices in system design. This heavily depends on the nature of the relationship and if it requires the employment of a classification algorithm in the feature selection layer of the forecasting system. If this is not the case and there is full freedom of choice in that layer, then even a specific basis might not influence the possible choices in system design.

This chapter shows the influence of the behavioural economic basis of the forecasting system on the design choices and how a limitation in feature selection trickles down throughout the system and restricts the use of methods in the following layers. As was described in the previous chapter, there exists very little knowledge about these connections between the layers. In the following section these inter-layer relationships and the restrictions that arise with the selection of a method in one layer on the decision freedom in the other layers is investigated.

6.2 Method choices in the framework's layers and their effect

In the previous section about the influence of the behavioural economic basis of the system and in the pre-analysis in chapter 4, it becomes clear that the methods selected in every layer of the system influence each other. In this section this relationship is investigated. To do so, first an enhanced overview of the methods for each layer including a grouping for similar methods is introduced. Then, each layer is reviewed successively as to how the selection of every method influences the following layers and if these relationships are similar for all methods within a certain group or if they differ from each other.

6.2.1 Framework overview including methods and their groups

In order to investigate the relationships between the layers, the current available framework overview that is shown in figure 4.2 is extended in a first step to further include the established methods and the group of methods that were introduced in chapter 4. The reason for this is to provide a better overview and to introduce the possibility to investigate if the influence of certain methods on further methods is the same for all methods within a group on another group of methods in the following layer.

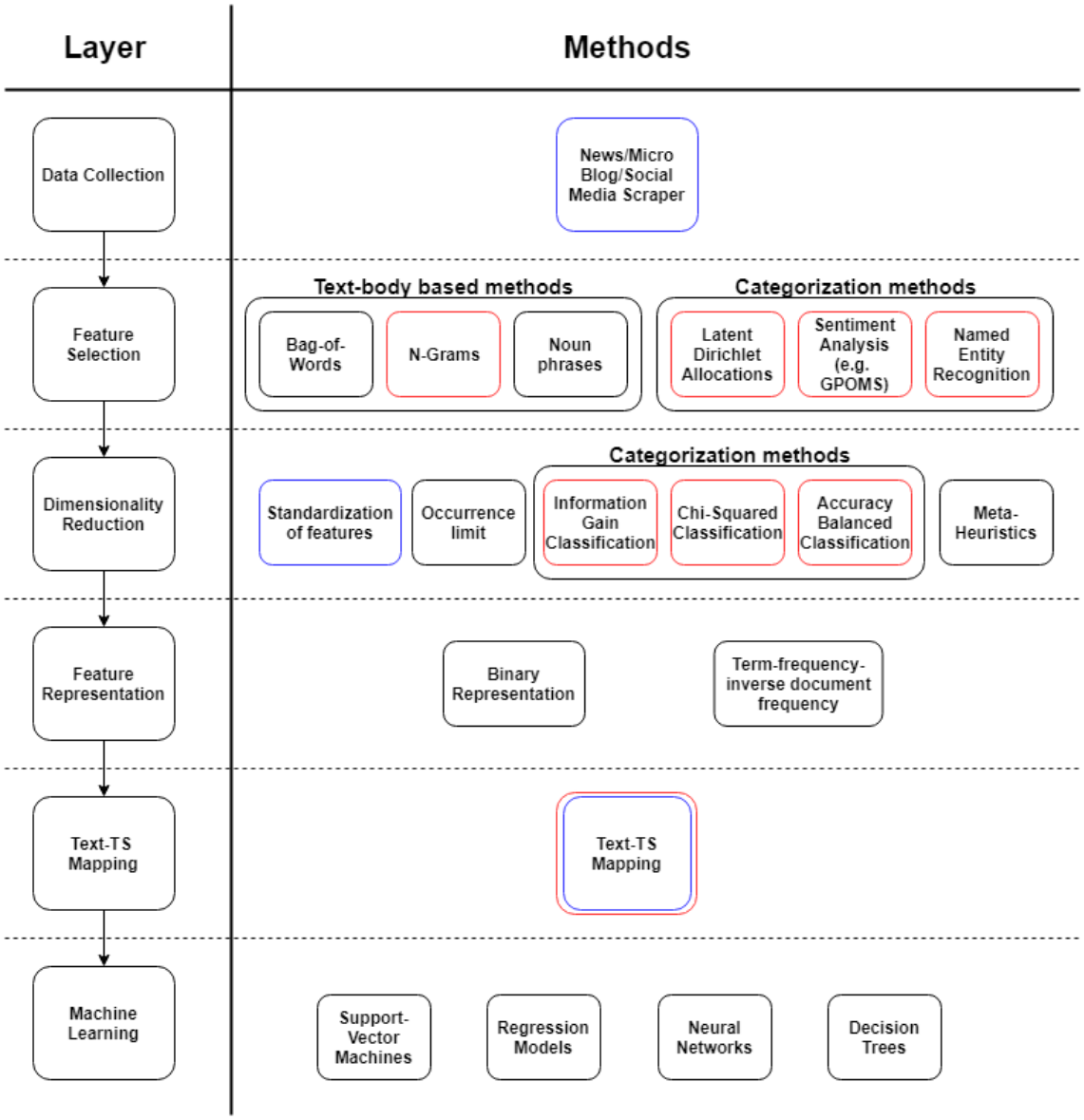

Figure 6.1: Framework including methods and grouping of similar methods. Red methods indicate that they need some pre-defined input data and blue methods indicate that these are mandatory steps that have to be conducted.

The enhanced framework is shown in figure 6.1. All methods that are marked in blue are methods that need to be used in any text mining based time-series forecasting system, regardless of the other system design. The methods that are shown in red are methods that require some form of pre-defined parameter input. These are mostly categorisation methods that require that the classes in which the input is supposed to be categorised are known in advance. The other two methods that require parameterisation are the N-Gram feature selection method, in which the number of words per N-Gram needs to be pre-defined, and the text - time-series mapping, where the lag needs to be determined beforehand. Based on this enhanced framework the relationships will be described in the following sections.

6.2.2 Relationships of the Data Collection layer

As mentioned above, naturally data collection is one of the steps that have to be conducted no matter what the rest of the system design is. In the section about the influence of the behavioural economic basis, it was established that the data that is collected is possibly influenced by underlying theories like the mood based market behaviour hypothesis, where it is imperative that only data that indicates the authors feelings is collected. Otherwise, the data collection layer does not have strongly relevant relationships with the other layers of the system, since the textual data that is scraped can be processed by all popular methods in the feature selection layer, no matter which of the methods is chosen.
Therefore it is fair to say that the data collection is influenced by the behavioural economic basis but does not hold any influence or restrictions over any other part of the text mining based time-series forecasting system.

6.2.3 Relationships of the Feature Selection layer

The next layer in the system is the feature selection, which lies between the data collection and dimensionality reduction layers. As described above the previous layer, the data collection, does not have any particular influence on the methods that can be chosen here. While the data collection might not have a strong impact on the feature selection, the feature selection can have quite some influence on data collection. This is especially the case when sentiment analysis or named entity recognition is employed. This relationship is based on the fact that those two methods are looking for very special attributes in the text, for sentiment analysis those are phrases like "I feel", "I am", "makes me" etc. [Johan Bollen, 2013] and for Named Entity Recognition it is the entities that are pre-defined, for example in a system that seeks to forecast Apple stock prices "IPhone", "Mac" etc. [Vu Tien-Thanh, 2012]. When employing these methods, it has to be ensured that input data texts are gathered that correspond with those requirements.
Anyhow, the analysis of the influence of the behavioural economic basis shows that it can have major implications on the degree of freedom in model choice here. A very specific basis theory can limit the methods drastically, which is the case for the mood based market behaviour hypothesis that only allows the usage of sentiment analysis in this step. In general, all future behavioural economic theories that are going to be used which refer to emotion as the main driver for predicted variables behaviour, regardless of what that predicted variable might be, arise the same restriction to sentiment analysis methods.
Looking at the influence that the individual methods in feature selection have on the successive layers, the following can be observed:

- **Bag-of-Words:** Bag-of-Words can be deemed the most simplistic method, where every word of a text document is simply seen as a feature. This simplicity is also the reason that choosing this method basically has no influence on which methods for dimensionality reduction can be chosen. Anyhow, since this approach leads to a large number of features, it increases the importance of the successive layer.
- **N-Grams:** It is shown in chapter 4, N-grams are very similar to Bag-of-Words, but instead of individual words groups of n words are selected. This similarity also can be seen in the influence on dimensionality reduction, where the selection of this method doesn't lead to any restrictions whatsoever.
- **Noun-Phrases:** The use of Noun-Phrases is slightly more sophisticated than the first two (although that does not indicate that it provides better results), where only the nouns and words related to them are chosen as features. While this also leads to no restrictions in the next layer, it has to be noted that employing this method

already leads to a significantly lower number of features to begin with compared to Bag-of-Words and N-Grams since the number of features is decoupled from the number of words in the text itself, which should be considered in dimensionality reduction.

- **Latent Dirichlet Allocation:** As described in chapter 4, Latent Dirichlet Allocation is a classifier that uses a pre-defined set of categories and assigns the articles to them. This method has serious implications on the dimensionality reduction since it basically makes it redundant to reduce the number of features and it makes it rather important that the significant features are selected. Since in the research's current framework no feedback loop between the machine learning algorithm and the dimensionality exists (which is analysed more in detail in the next section of the problem analysis), it becomes questionable if dimensionality reduction is necessary at all, since the number of features is already limited to the pre-defined list of classes.
- **Sentiment Analysis:** Because Sentiment Analysis also uses a pre-defined list of emotions or moods that are assigned to the text documents the same relationship exists between this method and the methods of the succesive layer as in Latent Dirichlet Allocation. Basically, the feature selection method already reduces the number of features to a rather small number, which renders dimensionality reduction almost redundant.
- **Named Entity Recognition:** Unsurprisingly, Named Entity Recognition also reduces the number of features to the number of pre-defined classes, so it limits what can be done in dimensionality reduction.

Based on the analysis above, it becomes obvious that the two groups of methods have the same implications for dimensionality reduction. Basically, the text-body based approaches hold no limitations for dimensionality reduction while the classification methods make the need for dimensionality reduction very questionable. This leads to the question if these methods should not be seen as hybrid feature selection/dimensionality reduction altogether, since they reduce the number of features to a set of pre-defined classes.
Further, the fact that in the current framework the standardisation of features is seen as part of the dimensionality reduction is debatable. While by definition, this process reduces the number features, it could be beneficial to do this already before the feature selection and directly after the data collection, since the obtained features for text-body based methods would be the same and the performance of categorisation methods might even be enhanced by providing standardised text as input.

6.2.4 Relationships of the Dimensionality Reduction layer and the Feature Representation layer

In the section above it is shown that the choice of method in the feature selection layer heavily influence the significance of the dimensionality reduction layer when classification methods are employed. Below the influence of the individual methods on the following layer, the feature representation, is described. Since there are only two different established methods in the feature representation, binary representation and term-frequency inversed document frequency, the relationship between each of the methods in dimensionality reduction and those two methods will be analysed. Since in some system designs, namely ones that do not employ any of the methods because they use classification methods in feature selection, the influence of that cases must also be considered.
In general, it must be said that no matter how the system is designed before the feature representation layer, binary representation is always applicable. Binary representation, which only considers the existence or absence of a feature, is applicable no matter if the

features are categories, words, phrases etc. A binary vector where the length is equal to the total number of features can always be constructed. Although this method does not rely on the previous choices in system design, this is not true for term-frequency-inverse document frequency. Like indicated in the name of the method here the frequency of a term inside a text document compared to the frequency in the whole input data is compared and returned as the representation of the feature. This naturally implies that it can only be used for terms, meaning words, phrases, groups of words etc. In conclusion, while this approach is feasible for all dimensionality reduction methods and therefore also for text-body based approaches in feature selection, it cannot handle categories, meaning that is not applicable when categorisation methods are employed in the feature selection layer.

6.2.5 Text - Time-Series mapping and machine learning

The text - time-series mapping and the machine learning layers differ from the previous layers in that they are completely independent from what happens before them. Since the time mapping part of the system solely cares about the time stamp of the individual input texts and does not consider their content whatsoever, it is completely independent. Similar to that, the machine learning part does not care about the content of the input data (as long as it is represented appropriately) but is rather focused on using in the training data to minimise some kind of forecasting error. It is important to mention that what is meant here is only that every machine learning methods usage is feasible, but the design of the system of course has an immense influence on the performance of the machine learning algorithm.
There is one special relationship between the text processing part of the system and the machine learning algorithm that lies within the text based nature of the system and the interpretability of text. The problem is that using text body based methods within the feature selection, especially bag-of-words but also N-grams and to a lesser degree noun phrases, provides features that in itself do not hold a lot of information. This can be shown in a very simple example.

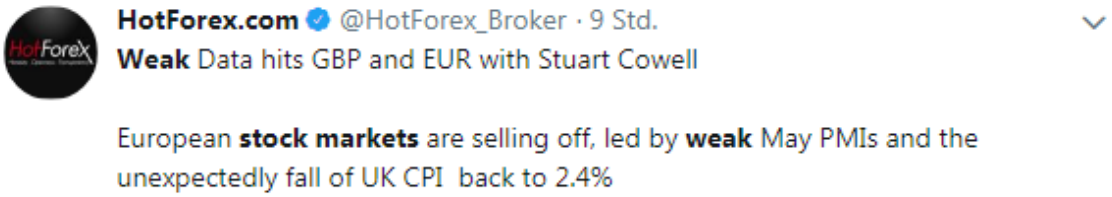

Figure 6.2: Example tweet for the importance of feature combination

Figure 6.2 shows a tweet that holds a lot of information, especially the sentence "European stock markets are selling off, led by weak May...". If that text is processed with a bag-of-words approach the individual words would be the features. That leads to the problem that observing the individual words in this example and in general holds very little information compared to the combination of words. While "stock markets are selling off" might be a strong signal for falling markets, only knowing the fact that all of these words appear in the text without any knowledge about the combination of them provides much less value. Therefore, a bag-of-words or a N-gram approach is much more promising when a deep learning method is used in the machine learning layer of the system rather than a supervised learning method, since these methods are capable of autonomously evaluating which feature combinations hold a strong predictive value, while linear models

like regression models are not able to "learn" from input feature combinations.

The analysis above shows that there is one main factor that influences the possibilities in system design significantly aside of the behavioural economic basis, the choice of categorisation methods in the feature selection method. No other decision in the design of the whole system has so many implications on the other layers as this one. Basically, the categorisation methods can be seen as feature selection/dimensionality reduction hybrid methods that make any further dimensionality reduction methods redundant and also limit the feature representation methods to binary representation.

6.3 The influence of the given forecasting problem

While so far the influence of the specific problem on system design has not been the subject of research yet, it still can't be ignored, especially when text mining based time-series forecasting is supposed to be used for real world problems. Surprisingly, it seems like there is very literature on that relationship in general, although the application of machine learning is dominating more and more of the daily life. Most of the arguments in this chapter seem self-explanatory, but it is important to understand the most obvious limitations that a specific problem might have on the design of a forecasting system based on text mining in order to apply such methods for real world problems.

6.3.1 Nature of the specific problem

It is self-evident that the nature of a specific problem has a large influence on the forecasting system. In this context, the meaning of the nature of the problem is the attributes like what is the predicted variable, how far ahead does the forecast need to look, is there a certain level of precision that needs to be achieved, what does the predicted variable depend on and so on. A simple example is the comparison of a machine learning algorithm that classifies medical pictures and one that is a short term intraday market forecast. While the first type of algorithm requires a large degree of accuracy, for the second type of algorithm it might be more important that it is able to deliver a relatively good forecast while having a minimal computation time. These performance criteria of the problem need to be considered when designing a forecasting system. Further, the architecture of the machine learning solution will be significantly different if the problem is to forecast something continuously or if the task is to create a one time forecast.

6.3.2 Computation speed

As was stated above one main criteria of a machine learning algorithm might be that it creates forecasts as fast as possible. This can have a significant influence on how often a machine learning solution is trained, since this training is rather computational expensive, especially for deep learning methods. This requirement is especially relevant for working with text, since the scraping of the text and the additional processing also needs to be accounted for when looking at computation speed. A good example that shows this problem is, again, the forecasting of intraday price movements. For market prediction it can be assumed that the earlier a forecast is available the more valuable it is, since this leaves more time to act upon it and therefore more possibilities to take

positions might occur in the market. This leads to certain requirements for all parts of the forecasting system. Such a problem requires that the data collection scrapes data live, and not in retrospective. Further, it would reduce computation time, if the unstructured input would also be processed and the result is stored live, instead of right before the forecast is created. This decouples the pre-processing and the machine learning part of the system chronologically. Further, the concept of lag, which was introduced in chapter 4 becomes more important, since it indicates the earliest point of time the forecast computation can start, while taking all relevant inputs into account. It is also necessary that the training of the machine learning algorithm is completed before that point of time to ensure a minimal computation time.

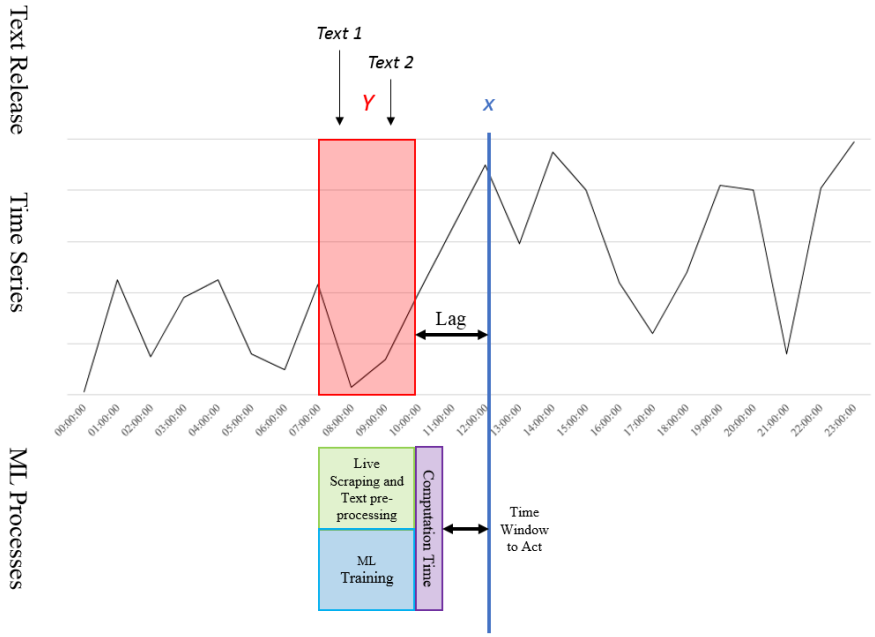

Figure 6.3: Example for the timing of ML processes in a system that reduces computation time

Figure 6.3 shows a example for a system that reduces computation time. The time window to act can easily be computed as $lag - computationtime$. In addition to the measures that are discussed above, it is beneficial to select methods in the machine learning part of the system that are faster than others.

6.3.3 Static nature of the environment

Another factor that needs to be considered in designing a forecasting system is, if the environment of the predicted variable is static. This has a major impact on the training of the machine learning model. Basically, this considers the question, if the relationship between the predictors and the predicted variable changes over time. In this case, it is not feasible to simply use the whole available historic data to train the model but time windows have to be identified, in which this relationship was similar to the current relationship. The impact of this effect on text mining based time-series varies, depending on the feature selection method. While the relationship between a small number of categories and the predicted variable can be modelled, the relationship between features obtained by a bag-of-words or N-gram approached cannot be easily tested, based on the interpretability issues of single words that was discussed in chapter 4. Consequentially, it might be beneficial

for systems of this nature to test, if a limitation of the training data to data points that are relatively close to the actual forecast increases performance, which would also be an indicator that the environment is not static.

6.4 Complexity of the system

In the pre-analysis in chapter 4 a short overview of the complexity of the field is provided, caused by the need to apply up to seven different methods including some pre-defined parameters. This fact, combined with the lack of research about how different combinations influence the forecasting performance, makes it almost impossible to find one best system configuration. Additionally, the lack of feedback loops in the current framework further means that in today's framework the optimisation of performance is actually not an objective. This whole issue that the system design in no way guarantees optimal performance can be seen as a "one-way paradox".

6.4.1 The one-way paradox

Text mining based time-series forecasting in its current form can be seen as a system with four overall steps, first the data is collected, then it is processed to make it readable for a machine learning algorithm, then this processed data is mapped on a time-series and finally this mapped data is used together with historical and other structured data to create a forecast.

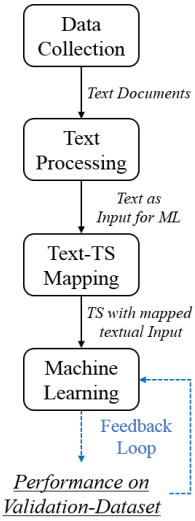

Figure 6.4: Overall System Process

This process is shown in figure 6.4, including the output/input data for every step. As the figure shows, the only part of the system that is subject to a feedback loop based on performance is the machine learning, where all methods have some mechanism included that optimises their performance based on a training data set. The rest of the system can be seen as static, where neither methods nor parameters are changed to increase the forecasting performance.
This static nature of the system is the one-way paradox that can be observed in all text mining based time-series forecasting systems. While the way the data must have a big

impact on the overall performance of the system, the first three steps of the process including their parameters are never subject to performance feedback. To illustrate this further, the whole system can be seen as an optimisation problem, in which the objective is to minimise the forecasting error in a given validation or training dataset based on some performance measure (e.g. MSE). While it would be intuitive to see as many of the parameters in the system, for example the lag in the text - time-series mapping or the minimum occurrence limit etc. as decision variables that are determined during the optimisation, in today's framework and approaches those variables are seen as constants that are pre-defined. In order to find an optimal system design, it would be necessary to go one step further and to even see the chosen methods in every step as decision variables. The reason why this has not been done so far is that it is close to impossible to create a system like that because it would be simply computational infeasible based on the sheer number of possible combinations of parameters and methods available. While it is not possible to create such a optimal system, it would be possible to include more parameters into the feedback loop, like the above-mentioned lag in the mapping process. The problem here is that the "higher" the feedback loop goes, meaning the earlier in the whole process decision variables get changed, the larger the impact is on the whole system.
The "one-way" paradox leads to one major insight: with the current knowledge it is impossible to create an optimal text mining time-series forecasting system, it is rather only possible to create working approaches, where a working approach is defined as a forecasting system that performs better than a forecast that does not incorporate the unstructured data.

6.5 Conclusion of the analysis

In this chapter the field of text mining based time-series forecasting has been analysed based on the problem statement, which was to investigate the influence of the behavioural economic basis on the system and the relationship between the different methods. The analysis conducted in this chapter concluded with the following findings regarding this research question:

- **Behavioural economic basis:** It is shown that the chosen behavioural economic basis has an influence on the system design that strongly depends on how specific the underlying theory is. While a very open hypothesis provides a large degree of freedom in the choices that can be made in system design, a very specific behavioural economic basis restricts which methods can be employed, but if it is proven that the hypothesis is true, it might lead to a higher chance of creating a successful forecasting system.
- **Method relationships:** The analysis argues that there is one major impact on system design based on which methods are chosen in which layer, and that is the selection of the group of methods in the feature selection. In contrast to categorisation methods, choosing a text-body based method here does not influence the choices in the successive layers. Choosing either of the presented categorisation methods has a very large impact on the rest of the system, since it basically makes dimensionality reduction redundant and forbids the use of other methods in the later stages.
- **Problem's nature:** While the analysis of the first two topics provide conclusive evidence that they influence the system design, this is not the case for the problem's nature. In general, it is valid to say that the specific problem has more influence on the requirements for the usage of forecasting system than the selection of the methods.

Additionally to the main research question, the influence of the complexity of the system was investigated. This analysis lead to the conclusion that based on the "one-way paradox" that lies within the nature of these forecasting system it is close to impossible to find a system that delivers optimal results.
In the following chapter, the solution part of the thesis, alternative approaches will be introduced that attempt to solve some of the problems that were found in this chapter.

7. Solution

While the pre-analysis in chapter 4 gives an introduction to the topic and shows the overall problem areas and the analysis in chapter 6 investigates those problem areas further, this chapter's purpose is to provide alternative approaches that provide a solution. In the following sections a new proposals is introduced, an adaption of the overall framework that incorporates the findings of chapter 6.

7.1 An enhanced system framework for text mining based time-series prediction

The previous chapter shows the influence of the behavioural economic basis and the decisions made in different layers of the system on the overall system design. The current framework shown in figure 4.2 does not incorporate these influences and can therefore be misleading in terms of the design of such a system. Neither does it acknowledge that the underlying hypothesis that the system is based on has any influence on the system, nor does it show the relationship between the single layers. When designing a new text mining based time-series forecasting system for a given problem, it might be used as a starting point or an indicator of what is needed, but it cannot be used as a guideline for the system architecture. Therefore, a new enhanced framework that fulfils the requirements needed is proposed below, to be used as such a guideline for creating a system architecture. This suggestion comes in the form of a flowchart, which is described in the subsection below.

7.1.1 Flowchart

The decision to use a flowchart to model is made for couple of reasons. First, the flowchart is a highly standardised method that is part of the ISO regulation which is widespread through a number of fields. Further it is easily readable and very flexible and provides a good level of detail of a process, which is the main reason that it was chosen in this thesis [Ruth Sara Aguilar-Savén, 2004].
A flowchart consists of multiple elements. Figure 7.1 explains the elements that are used in the new proposed framework.

Symbol	Name	Description
	Arrow	Connects the elements of the chart and indicates the flow of the system.
	Terminal	Indicate events that happen within the process. Are also used for the start- and endpoint of a process.
	Process	Indicate events that happen within the process. Are also used for the start- and endpoint of a system.
	Decision	Shows a condition that determines the following elements of the flow.

Figure 7.1: Elements of a flowchart diagram

This method is applied to illustrate the process of designing a new forecasting system based on a specific problem. Additionally to the fact that it is easily understandable, the flowchart method is chosen because it can be used as a template for constructing a system, without knowing much of the fundamentals of the field, since some choices only have to be made if a certain path is chosen.

7.1.2 Introduction of the new system construction framework

In this section, the system design framework, which is shown in 7.2, is introduced and described in detail. Before that, it needs to be stated that this new approach is aimed at real world problems, where a specific problem is the starting point, rather than for research problems, where the sequence of problem definition, data collection and finding the behavioural economic basis can be quite fuzzy. Further, it must be noted that in parts of the analysis it is discussed that in certain setups of the system some methods might perform better than others. This is not part of the new framework, because the underlying idea is to show the necessary steps based on which method is chosen. The whole system can be roughly separated into three parts, first the preparation phase that includes defining the "logical" basis of the system by defining the requirements of the problem and the behavioural economic basis. The second part is the "actual" system design process, in which the methods are chosen based on the findings of chapter 6. The last part is the prediction part of the system, in which the text is mapped and a machine learning model is employed and trained. In the following list, this system is described step by step, where processes are marked with a **"P"** and decisions with a **"D"**:

Start to feature selection:

- **P - Define Problem:** As mentioned above, the starting point of the system is a specific forecasting problem, be it to predict intraday prices of a certain market or the number of passengers on a certain train route. The first step is to define the requirements of this problem in a machine learning sense, as discussed in chapter 6. That means a time horizon has to be defined, a performance metric, what does the output of the system need to be etc.
- **P - Identify Behavioural Economic Basis:** The second step is defining the behavioural economic basis of the system and therefore establishing how text data is connected to the predicted variable. This can either be done by choosing an established theory, for example the adaptive market hypothesis, or by defining a new

hypothesis based on previous findings, similar to the work of Johan Bollen [2013].

- **P - Collect and standardise Data:** As soon as the behavioural economic basis is chosen the data selection process can start, in which textual input is selected that corresponds to the decisions in the previous step. In this framework, this step also contains the feature standardisation. Given that this process is seen as part of the dimensionality reduction in the classical framework, this might seem questionable at first glance, but there are valid reasons to conduct this step before the feature selection method. First, doing it before the feature selection prevents that features are selected that are going to be deleted in a next step anyways, if a text-body based approach is chosen. Second, it does not make sense to do this after a classification method is applied, if a classification approach is chosen, since then the input documents are already reduced to their classes. Further, it is standard practice to standardise the text before it is classified [Kumar Ravi, 2015], therefore this is step is moved above the feature selection layer.
- **D - Behavioural Economics based on emotion:** At this point, the system splits for the first time based on the behavioural economic basis. As discussed in chapter 6, the choices in system design are reduced if mood or emotion are involved in the feature selection layer since this means that only sentiment classification methods are feasible.
- **P - Choose Feature Selection Method:** If no behavioural economic basis is selected that relies on emotion, the next step is to choose a feature selection method. At this point all methods presented in chapter 4 are feasible.
- **D - Classification Method Chosen:** If a classification method should be chosen in the feature selection layer, the rest of the process becomes the same as for systems that are based on emotions. Else, the process continues in the text body based branch of the system (shown on the left side in figure 7.2).

Text body based branch:

- **P - Employ Feature Selection Method:** The feature selection of the (already standardised) features is conducted.
- **P - Choose Dimensionality Reduction Method:** In this branch of the system it is imperative to conduct a dimensionality reduction, where all methods are feasible. Anyhow, the number of features that are subject to this reduction depend on the chosen feature selection method, as the number of features that needs to be deducted is bigger for bag-of-words and N-grams compared to the result of a noun phrases method.
- **P - Define Parameters:** In this step, the parameters for the rest of the system need to be determined. This greatly depends on what method was chosen in dimensionality reduction and the problem itself, because the lag and the relevant time window for the text-time series mapping are also selected here. This is important for being able to split up the dataset into smaller pieces that are easier to handle from a computational point of view. When setting the parameters for the dimensionality reduction a threshold (number of observations, minimum occurrence limit etc.) is introduced and all features that don't reach it are deleted.
- **P - Employ Dimensionality Reduction Method:** Here, the selected method is conducted, employing the defined parameters.
- **P - Choose and Employ Feature Representation:** After the features are reduced to the target level, the last step in the pre-processing is to make them readable for the machine learning algorithm. This is done in this step, where either binary representation or term-frequency-inverse document frequency is chosen and employed immediately.

Classification branch:

- **P - Define classes:** If no pre-defined classes are available, which isn't the case in most sentiment analysis approaches, then the relevant classes have to be defined in this step.
- **P - Employ Classification Method:** In this step the classification of the documents is carried out.
- **P - Employ Binary Representation:** In chapter 6 it is discussed that of the two available feature representation methods, only binary representation is applicable in this branch (since a document is either a member of a class or not), therefore this method is employed.

Text - Time-Series Mapping to End:

- **P - Conduct Text - Time-Series Mapping:** In the "Define Parameters" process above, the lag and the size of the relevant time window have been defined and here these parameters are used to map the processed text documents against the time series.
- **P - Choose Machine Learning Method:** Here, the machine learning method is selected. While it is feasible to select any method regardless of the rest of the design of the system in theory, chapter 6 shows that it is more promising to use deep learning methods (e.g. decision trees and neural networks), if the system employs a text body based approach, since they are able to establish relationships between the features autonomously.
- **Employ Machine Learning Method:** Finally, the machine learning method is used to create a forecast. This process also includes the optimisation of the machine learning algorithm, where the parameters are tuned to ensure a good performance. Which parameters are there to tune depends on the selected method.

Using this flowchart to construct a text mining based machine-learning model helps to create robust forecasting systems rather easily, since it shows which methods in every layer are compatible. Further it incorporates the influence of the behavioural economic basis of the system, which has a major impact on the design as is elaborated in chapter 6. Anyhow, employing this approach in no way guarantees that the forecasting system will perform well, since the analysis of the complexity of the system shows that every choice in the system construction, be it methods or parameters, influences the performance in the end. How this problem could be approached is part of the discussion chapter in this thesis.
In the next section an example is shown for how this framework can be used to create such a forecasting system based on real world data.

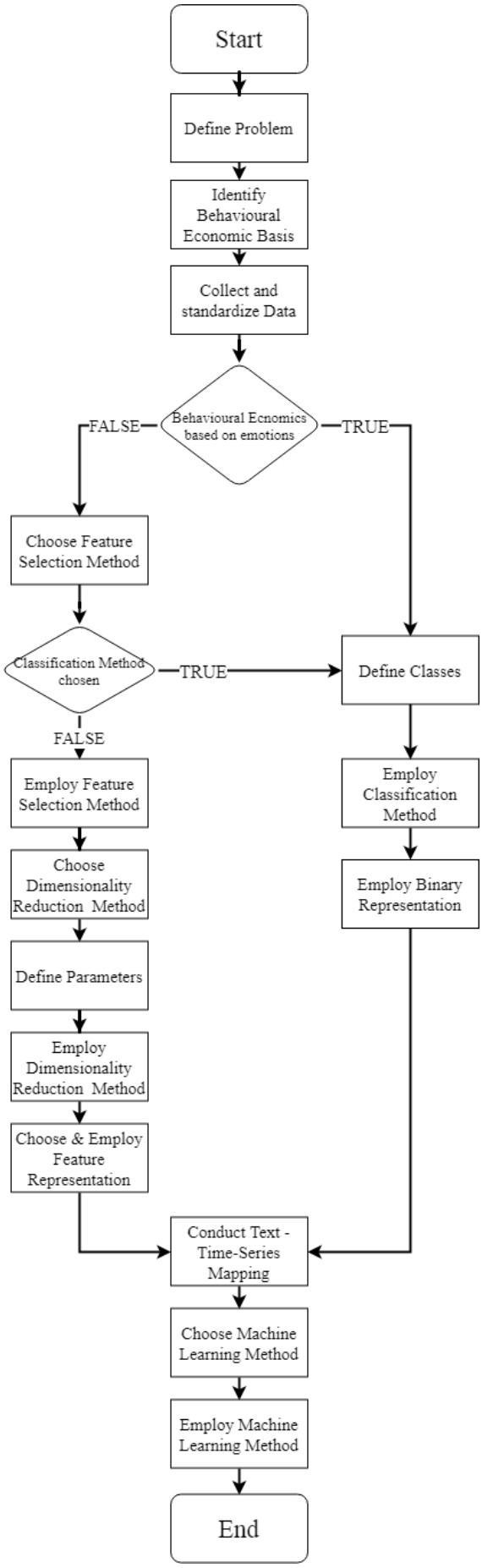

Figure 7.2: A new approach for creating a text mining based time-series forecasting system

7.2 A real world example of the framework application

In this section of the thesis, the framework that was developed is employed to create a text mining based time-series forecasting system. The problem that to forecast the intraday (ID) closing prices of the Dow Jones Index (DJI) for a 30 minute time interval was chosen for this example. The selection of this forecasting problem is based on multiple reasons. First, while it would have been interesting to venture in a field outside of financial markets, the fact that the whole research body is based on market prediction makes it easier to stay within that topic for an simple example. Second, financial markets are quite a transparent field, where it is comparatively easy to obtain data. Another reason for the selection of this example is, that while it is rather easy to scrape live data, getting historical data (for both the predicted variable and the text data) can be quite hard and possibly very expensive. Therefore instead of forecasting daily prices, which would have resulted in a very small dataset, a higher resolution is used. The rest of this section is structured in the same way as the flowchart presented above.

Figure 7.3: Chart of the 30 minute ID prices of the DJI from Thomson Reuters

7.2.1 Define Problem

The first step is to define the problem in terms of what the expectation of the machine learning algorithm are. In this case, while the desired information is if the markets are going up or down, it is not relevant to know exactly how much the markets are going to move. Based on this, it is feasible to employ a binary classifier as a machine learning algorithm, which is also the forecasting approach that received most attention in research [Arman Khadjeh Nassirtoussi, 2014]. Additionally, a baseline has to be defined in this part of the framework. For this problem this baseline to beat is a naive forecast, which is a random walk that would lead to a accuracy of 50% in the long run. This approach is also widely spread in literature and is for example used in the work of Michael Hagenau [2012]. Since this experiment would be seen as a backtest of a forecasting system, rather than a live system in a real world application, there are no pre-requistes regarding computation time. Additionally, one hypothesis needs to be made beforehand, and that is that the relationship between the selected textual input and the intraday price of the DJI is static.

7.2.2 Identify Behavioural Economic Basis

In this example, the adaptive market hypothesis is chosen as the behavioural economic basis of the system. This theory implies that the markets are not efficient and therefore can be forecasted and that the individuals that compete in the market are learning based on the information they consume. Further, the openness of the adaptive market hypothesis allows the use of a wide range of text sources, under the requirement that market participants use them. This link between the textual input and the market prices ensures that the use of a text mining based time-series forecasting system is valid.

7.2.3 Collect and standardise Data

The next step in the system is to identify, scrape and standardise the textual data. In this example tweets are selected as the data source. The usage of these micro-blog entries has a couple of advantages. Since the forecasting time window is rather short based on the problem (30 minutes), it is important to ensure that enough input data is available for each observation. Using twitter as a data source ensures that, because new tweets are constantly created. Another benefit of using tweets is, that both news sources use it to publish links to their articles and market participants use it to voice their opinion. The danger here is that this information could get lost in the sheer number of tweets created. The data was scraped from the twitter search API, which is publicly available but also has the limitations that only data of the last seven days is available and that the number of requests that can be sent to the API is strictly limited. Therefore, it was not possible to simply scrape all tweets that were created in the observation time window, because there are more tweets created than can be scraped. It was necessary to introduce the following limitations:

- Only select tweets that are written in English.
- Only select tweets that were published in North America.
- Limit the scraper to tweets that contain certain search criteria, based on a set of thirty six terms. The phrases that were selected are based on the work of Tobias Preis [2013], who establishes a connection between certain search terms on Google and the movements of the DJI.

The scraper ran over the course of three months and collected roughly 17.5 million tweets in total. As a source for DJI prices Thomson Reuter's Eikon platform was used. The data was collected for the timespan from 26-03-2018 to 18-05-2018 and only the official trading hours where selected, which leads to 580 observations in total.
After the data was collected, the text was cleaned and standardised. This was done using the standard procedure described in chapter 4 and additionally, some further measures were necessary based on the special attributes of tweet texts. Since twitter allows the use of emoticons, which can not be processed by text mining techniques, all words that contain characters that are not compliant with the UTF-8 standard were deleted. The same was done for all words containing "http", which indicates weblinks and cannot be used, because of their lack of interpretative value.

7.2.4 Choose and Employ Feature Selection Method

Since the forecasting system is not based on a emotions, in the next step of the process the feature selection method has to be chosen. In this example bag-of-words was selected, based

on the fact that it is the predominant technique used in literature [Arman Khadjeh Nassirtoussi, 2014] and that this technique is the simplest one to apply. This decision implies that in the example presented in this section the text body based branch of the framework is carried out, which was done on purpose since it requires more methods and therefore provides a better overview. Applying bag-of-words on the collected tweets delivers a total number of around 1.7 million features, where the most frequent one is "dow" followed by "jones".

7.2.5 Choose Dimensionality Reduction Method and Define Parameters

In this part of the system the simplest method was chosen, minimum occurrence limit, since the idea behind this example is to show the process rather than to create the most complex, best performing system. Because of technical limitations it was necessary to cut down the number of features quite drastically. The occurrence limit was set to a 100,000, which leads to a total number of 243 features. The main reason for that cut is, that an integer uses 32 bits of storage in the programming environment that was used. This means that for every feature selected 68 megabytes of storage are required (given that there are 17 million tweets), which makes it unfeasible to select a large number of features on a regular computer. In addition to the occurrence limit the parameters for the text - time-series mapping must be defined in this step. For this experiment a time window of the length 30 minutes was chosen to ensure that the same textual data is not used for two different observations. The lag was chosen to be 30 minutes as well, based on the findings of Tarun Chordia [2005], who shows that the time span that it takes for an inefficient market to converge to efficiency is somewhere between 5 and 60 minutes. That means that the market needs time somewhere in this span to absorb the information and to include it in the price.

7.2.6 Employ Dimensionality Reduction Method and Choose and Employ Feature Representation

Using the parameters defined in the previous process, the minimum occurrence limit technique is employed. Next, the feature representation was selected and carried out. In this framework, binary representation was chosen, since it is computationally cheaper than it's alternative and there is no proof that it performs worse. The result of this step is the matrix that is shown in figure 7.4, including all available tweets with a timestamp and all 243 features. This is the final step of the text pre-processing and the only process that is left before it can be fed into a machine learning algorithm is to map it to the time-series.

tweet nr	timestamp	also	alway	amazon	american	amp	analysi	anoth	app	appl	april	ask
1479	26-03-2018 14:37	0	0	0	0	1	0	0	0	0	0	0
1965	26-03-2018 14:37	0	0	0	0	1	0	0	0	0	0	0
2803	26-03-2018 14:38	0	0	0	0	0	0	0	0	0	0	0

Figure 7.4: Example for the result of the feature representation

7.2.7 Text - Time-Series Mapping and Selection of the Machine Learning Method

The mapping of the processed text is the last step that needs to be conducted before the machine learning part. This can easily be done using the parameters that were defined beforehand and the timestamps in the matrix shown in figure 7.4. For every observation of

the predicted variable the data simply has to be filtered for tweets that were published between 30 and 60 minutes before the observation.
After this is completed, a machine learning method has to be chosen. Since bag-of-words is used in the feature selection method, it is recommended to use either decision trees or neural networks as described in chapter 6. Since there is no way of evaluating the method that will perform the best beforehand, both methods were tested and the more promising algorithm was chosen. But before that could be done, the dataset of 580 observations needs to be split up into training data to train the method, validation data to perform the parameterisation of the method and test data to evaluate the final performance. It is important for the robustness of the performance analysis that the dataset is split up in three sets rather than only training and validation data, since the goal of machine learning is to do well for data the method has not seen before [Francois Chollet, 2018]. Consequentially, the machine learning process is the following: the model is trained on a training dataset, so it performs optimally given a certain parameter configuration. Once the model is finished learning, it is used in the validation dataset and the result is stored. This is done for a set of parameter configurations with the goal to maximise the accuracy in the validation data. Since the validation data is used to optimise the model's parameters it must be deemed as data the model has seen before, so using the accuracy of the best performing set of parameters in the validation data is not feasible. This is why the model with best performing parameter combination is then applied to the test dataset and the accuracy that results from this test is seen as the fitness of the model.

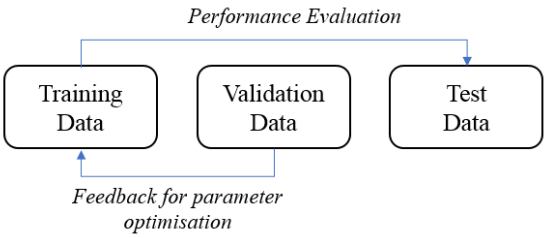

Figure 7.5: Relationship between training, validation and test data

In this experiment, the split 80% of the data is used for training and 10% for validation and testing respectively. Rather than selecting random samples for each set, the data was split chronologically. This was done on purpose, since a bad performance in the validation data set would indicate that the hypothesis that the relationship between tweets and DJI ID prices is static is wrong.
For the initial test a sequential neural network with three densely-connected hidden layers and two dropout layers to reduce overfitting has been constructed. This was tested with several different parameter combination, which all have shown the same result. The model started overfitting the training data very fast, even with high dropout rates, and the performance on the validation data was rather bad, as can be seen in figure 7.6, where the upper part shows the loss for each iteration and the lower part the accuracy.

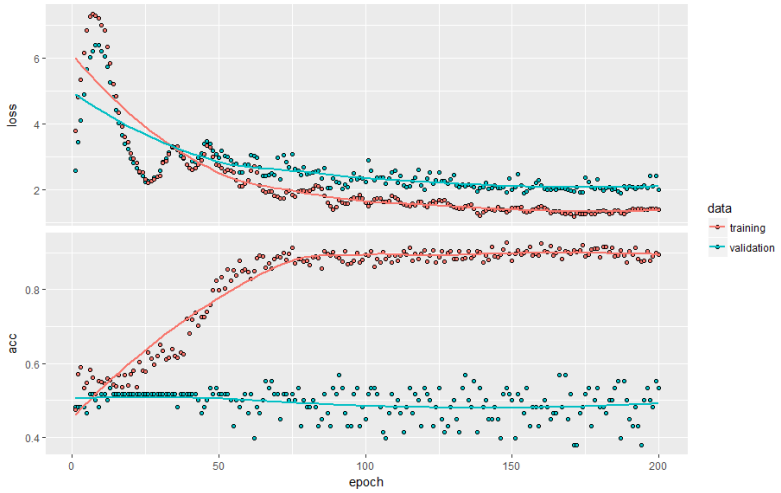

Figure 7.6: Performance of a neural network with three hidden layers on training and validation data

Based on these results, the decision was made to employ a tree based approach as the machine learning algorithm.

7.2.8 Employ Machine Learning Method

In the last step of the framework, the decision tree method was applied, parameterised and finally tested. To ensure a good performance, an enhanced version of the tree based algorithm was used, called boosting. The idea of boosting is to not only fit one tree to the dataset, but to fit a lot of small trees consequentially. This is done by iteratively fitting a tree to the residuals of the current forecasting model and then updating the model with the new tree [Gareth James, 2013]. These models have multiple parameters that can be used to optimise performance:

- The number of trees used
- The shrinkage parameter, also referred to as the learning rate
- The number of splits in every tree
- The dropout rate for both observations and/or features, which is used to prevent overfitting

For this experiment, the parameter optimisation has been conducted with a brute force approach, meaning that a number of pre-defined parameter combinations have been tested on the validation data and the best one is selected. To create this set of parameter combinations the approach suggested by Owen Zhang [2015] has been used, who suggests to keep the number of trees fixed between 100 and 1000 and perform a grid search for the other three parameters. Since initial tests have shown that a large number of trees lead to overfitting, the number of trees has been fixed at 100. For the other four parameters, three levels were pre-defined each, which leads to 81 combinations in total, and tested. Figure 7.7 shows the parameter levels that were tested.

Learning Rate	0.004	0.01	0.02
Number of Splits	8	10	12
Feature Subsample	0.25	0.5	0.75
Row Subsample	0.2	0.4	0.6

Figure 7.7: Tested parameter levels

Employing the brute force approach on all possible combinations of this set shows that the best combination, measured on the accuracy in the validation set, is to set the shrinkage to 0.02, the number of splits to 8, the observation dropout rate to 25% and the feature dropout rate to 75%. This parameter configuration leads to an accuracy of 60.34% in the validation set and 58.62% in the test set. Since the goal of the whole system was to beat a baseline of 50%, this text mining based time-series forecasting system can be viewed as a working forecast.
Anyhow, while the accuracy for in the validation set for the different parameter combination varies between 36.20% and 60.34%, the accuracy in the training set is above 80% for all configurations. This is a clear indicator that the machine learning algorithm is overfitting. The reasons and implications of this overfitting are elaborated in the next chapter of the thesis, the discussion.

8. Discussion

So far this thesis provided an overview of the field of text mining based time-series forecasting, the problems in the current research body were analysed and a new framework was developed that can be used to create such a forecasting system. Additionally, a real world example was introduced to show how that new approach can be employed. This part is discussing the findings that were presented so far. The first topic is the performance that was achieved in the experiment. Another topic that is up for discussion is the universality of the newly presented framework and what could be improved. After that, the possibility to create a system that is optimised holistically is investigated, based on the analysis of the complexity of text mining based time-series problems that is shown in chapter 4 and 6. Further, future research that would be beneficial for the field is discussed.

8.1 Overfitting in the experiment

While the experiment in chapter 7 shows results that beat the baseline for a binary classification problem, the performance that is achieved is still debatable. The main reason for that is, that both the neural networks and the decision trees model are performing very well in the training data, with accuracies in the range between 80% and 90%, but compared to that performs poorly in the validation and test data, although measures like dropout were introduced to prevent that. This effect is called overfitting of a machine learning model and is a common problem in both research and practice, since it decouples the performance in the training dataset from the performance that is achieved on data that the model has never seen before. There are multiple possible reasons that this happens in the example of forecasting ID DJI prices:

- **The size of the dataset:** While on first glance it might not look very small, 580 observations are actually considered a small number of samples in a machine learning context [Francois Chollet, 2018]. One reason for the overfitting that we see in the experiment's results could simply be that there is not enough data to train on, and therefore the machine learning model sees patterns in the provided training data that are not universally valid.
- **The relationship between the text data and the ID DJI price changes over time:** As discussed in the problem definition process of the experiment, the hypothesis that patterns in the input data set have a constant influence on the predicted variable over time. One of the reasons for the difference in performance in the datasets is that this hypothesis does not hold true. It is very likely that the market drivers change over time, while in the beginning of observed time-series some policy decisions might have been the main motivation for market movements, in the last 20% something completely different could influence the prices. This theory is also somewhat supported by the fact that the performance in the test dataset, which is chronologically further away from the training dataset than the validation dataset, is slightly poorer.

- **Sub-optimal parameterisation of the system:** Like Hutson [2018] claims, machine learning has a certain feel of alchemy and this is especially true for the paramterisation of deep learning models, like the ones employed in the experiment. While some guidelines in how the parameters are supposed to be set can be found, the performance of certain combinations is very problem specific. Even though measures were taken that go beyond what is recommended to counter overfitting, it still is possible that the selected parameter combinations were simply not the correct ones.

All these reasons are possible explanations for the overfitting that can be observed in the experiment. Anyhow, it needs to be noted that the whole point of the experiment was not to create the perfect forecasting system, but rather to provide an idea of how the new framework can be applied to a real world problem.

8.2 Universality of the new framework

It is important to discuss what the newly presented approach can, and maybe even more relevant, cannot do. The whole idea behind the suggested extension to the framework is to translate the findings of the analysis in chapter 6 into a system that can be used to design a valid text mining based time-series forecasting system that takes all limitations provided by the behavioural economic basis and the decisions made in different layers into account. It is fair to say that this was accomplished for the current state of research, but it must be recognised that the literature body today is not very extensive. By default, using the today's research comes with certain limitations that apply to the system. First, there are only two established behavioural economic hypothesis that were used so far. It is very much possible that the introduction of a different theory will require the new framework to be updated. This goes hand in hand with the fact, that currently all research is done on market prediction. The application of text mining based market-prediction in new fields could also arise the need to update the framework. The same would be the case if new methods are introduced in the different layers.
In general, the new approach to design such forecasting systems cannot be seen as a static, universal truth, that can be used in it's current form from now on, but as the whole field evolves, which is more than likely since it is such a young research topic, the framework needs to evolve with it.

8.3 Holistic optimisation of the forecasting system

While the main topic of the thesis is to identify restrictions in the system and relationships between the different elements of the forecasting system, at some points it is also discussed which methods not only are feasible to used with each other, but also show more promise performance wise. One example for this is that it is recommended to use text body based feature selection methods together with deep learning methods, since they can "learn" which combination of features matter.
Having this in mind and conducting the method selection in every step of the experiment, the feeling arises that more educated guesses are taken than a sound decisions. Based on the lack of fundamental research, the question: "Is there a way to optimise a text mining based time-series forecasting system holistically?" arises. The simple answer is no. The more complex answer is maybe, one day. The reason for that lies in the complexity of the system, which was discussed in chapters 4 and 6. A simple example that shows how difficult such a optimisation would be, is just to consider the lag in the text - time-series mapping

step as a decision parameter. Instead of using a single, fixed value, the lag would need to be chosen dynamically. For example, the parameters in the experiment could be extended with the lag, which would also hold three different values. It can be easily computed how many parameter combinations that delivers by using the formula $levels^{parameters}$, which would increases the number of parameter combinations in the example from 81 to 243. Assuming three levels per parameter are not sufficient, but rather five should be tested, would lead to a total number 3125 combinations, and five levels per parameter is not a very large number. To make it even more complex, it needs to be considered that in text mining based time-series forecasting systems it is assumed that all that all feature combinations have the same lag, which is not the case in reality. The text *"the market will go up"* and *"the market did go up"* share all the same features, except for one word, which changes the whole meaning of the sentence. If the lag parameter is supposed to be truly optimised, it is necessary to consider these features that change the lag of the input as well.
This small example shows, how difficult it would be to only include one more parameter in the feedback loop. But in a truly optimised system both the parameters (e.g. lag, time-window, method-specific parameters like minimum occurrence limit etc.) and the chosen methods would have to be optimised, which goes beyond what is possible today.

8.4 Future Work

Based on the fact that text mining based time-series prediction is such a young field, which does not have a lot of literature published on it yet, there are some interesting aspects that definitely deserve to receive more research attention. The most promising future research topics are presented below.

- **Appliance in other fields then market prediction:** One of the first topics that come to mind is the appliance of the technique to different fields. While there are valid reasons that so far market prediction has received all of the attention, as discussed in chapter 4, there are no causes that it should not be possible to use the same approach for other forecasting problems. This goes hand in hand with either finding or developing new behavioural economic theories, that link textual data to time-series data.
- **Developing a deeper understanding of the different method combinations:** While this thesis mainly focuses on the feasibility of different method combinations, it would be beneficial to investigate the performance of different combinations as well. A first attempt to do so can be found in the work of Michael Hagenau [2012], but it is limited to using different methods for dimensionality reduction.
- **Developing more sophisticated methods across the framework's layers:** Many of the methods presented in chapter 4 are arguably quite simplistic, especially text body based feature selection methods. It could be very interesting to develop more sophisticated methods, that focus more on the parts of a text that contain the core message, for example adjective-noun or noun-verb combinations. Such more elaborate approaches in the feature selection layer would most likely also lead to the possibility to create and employ more sophisticated methods in the dimensionality reduction layer.

9. Conclusion

This thesis starts with the initiating idea of providing an explanation of text mining based time-series forecasting systems, finding an area which was not addressed in research so far and providing a deeper understanding of this area. To do so, this forecasting approach is described in detail and then analysed thoroughly and certain problem areas were identified.

As a result of this so-called pre-analysis in chapter 4, the research problem is stated focusing on analysing the factors that influence the design process of text mining based time-series forecasting systems. Additionally, the question about the overall influence of the complexity of these systems was addressed. Looking back at those investigated areas, the following conclusions are drawn:

Which factors influence the design of a text mining based time-series forecasting problem?
The analysis in chapter 6 shows, that there are two main influencing factors on the design of a text based forecasting system, the behavioural economic basis and the chosen feature selection approach. The first influencing factor, behavioural economics, is important, because some theories only allow the use of a classification based approach, while others provide a higher degree of freedom within the design process. The second factor, the choice of a feature selection method, indicates that the design of the system changes, if a categorisation method is chosen. These two insights lead to the conclusion that there are two different approaches when designing a text mining based time-series forecasting system, a classification and a text body based approach.

How does the complexity of the system influence the forecasting system?
In addition to the main research question, a second, subsidiary problem was introduced, the impact of the overall complexity of such forecasting systems. The analysis shows that, while this factor leads to some method combinations that make more sense than others, there are no definite limitations on the design process that arise from it. Nevertheless, the complexity of these forecasting systems leads to one problem, that it is very hard to optimise them holistically. This is also the main reason why the parameterisation is only done on the machine learning part, instead of the whole system.

These findings open the door for future research. One topic that needs to be investigated in more detail is the impact of the different feasible method combinations on the forecasting performance. Further, it needs to be determined how these systems can be applied outside of market prediction, finding new or identifying existing behavioural economic theories that connect text and time-series data. All in all the fact that this research field is still relatively new, but nevertheless was able to show promising results in market prediction makes it very likely that it will move more and more into the spotlight of research.

Bibliography

Aalborg University PBL Academy (2017). Problembaseret læring (pbl) på aalborg universitet. `http://www.en.aau.dk/about-aau/aalborg-model-problem-based-learning`. [Online; accessed 17-May-2018].

Alex J. Smola, B. S. (2004). A tutorial on support vector regression. *Statistics and Computing*.

Alexander Pak, P. P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Alper Kursat Uysal, S. G. (2014). The impact of preprocessing on text classification. *Information Processing Management*.

Andrew Urquhart, R. H. (2010). Efficient or adaptive markets? evidence from major stock markets using very long run historic data. *Journal of Computational Science*.

Andrew W. Lo, . (2005). Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. *Journal of Investment Consulting, Vol. 7, No. 2, pp. 21-44, 2005*.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, T. Y. W. D. C. L. N. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, T. Y. W. D. C. L. N. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*.

Azadeh Nikfarjam, Ehsan Emadzadeh, S. M. (2010). Text mining approaches for stock market prediction. *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*.

Burton G. Malkiel, . (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives Vol.17 No.1 Winter 2003*.

Burton G. Malkiel, . (2005). Reflections on the efficient market hypothesis: 30 years later. *The Financial Review 40 (2005) 1-9*.

Chih-Fong Tsai, William Eberle, C.-Y. C. (2013). Genetic algorithms in feature and instance selection. *Knowledge-Based Systems*.

Christopher J. Neely, P. A. W. and Ulrich., J. M. (2009). The adaptive markets hypothesis: Evidence from the foreign exchange market. *Journal of Financial and Quantitative Analysis, Vol. 44, Issue 2*.

D. Sculley, Jasper Snoek, A. W.-A. R. (2018). Winner's curse? on pace, progress, and empirical rigor. *ICLR 2018 Workshop Submission*.

Fang Jin, Nathan Self, P. S.-P. B. W. W. N. R. (2013). Forex-foreteller: currency trend modeling using news articles. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Francois Chollet, J. A. (2018). *Deep Learning with R*. Manning, first edition.

Gareth James, Daniela Witten, T. H.-R. T. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, eighth edition.

George Orwell, . (1949). 1984.

Hutson, M. (2018 (accessed May 15, 2018)). *AI researchers allege that machine learning is alchemy*.

Johan Bollen, Huina Mao, X.-J. Z. (2013). Twitter mood predicts the stock market. *International Review of Financial Analysis*.

John Gantz, D. R. (2013). Big data, bigger digital shadows, and biggest growth in the far east. Online publication by IDC.

John R. Nofsinger, . (2005). Social mood and financial economics. *Journal of Behavioral Finance, Volume 6*.

Kumar Ravi, V. R. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, Vol. 89*.

Matthew Butler, V. K. (2009). Financial forecasting using character n-gram analysis and readability scores of annual reports. *Canadian Conference on Artificial Intelligence - AI 2009: Advances in Artificial Intelligence*.

Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, M. E. B. (2009). Text feature selection using ant colony optimization. *Expert Systens with Applications*.

Michael Hagenau, Michael Liebmann, M. H. D. N. (2012). Automated news reading: Stock price prediction based on financial news using context-specific features. *HICSS '12 Proceedings of the 2012 45th Hawaii International Conference on System Sciences*.

Owen Zhang, . (2015). Winning data science competitions. Meetup event hosted by NYC Open Data Meetup, NYC Data Science Academy.

Paul C. Tetlock, Maytal Saar-Tsechansky, S. M. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*.

Peter Nielsen, Z. M. (2018). The impact of stochastic lead times on the bullwhip effect - an empirical insight. *Management and Production Engineering Review, Vol. 9*.

Robert P. Schumaker, H. C. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*.

Ruth Sara Aguilar-Savén, . (2004). Business process modelling: Review and framework. *International Journal of Production Economics Vol.90 Issue 2*.

Sanjiv R. Das, M. Y. C. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*.

Sebastian Mallaby, . (2010). *More Money Than God: Hedge Funds and the Making of a New Elite*. Penguin, first edition.

Serafettin Tasci, T. G. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*.

Sholom M. Weiss, Nitin Indurkhya, T. Z. F. J. D. (2010). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, first edition.

Tarun Chordia, Richard Roll, A. S. (2005). Evidence on the speed of convergence to market efficiency. *Journal of Financial Economics, Vol. 76*.

Tobias Preis, Helen Susannah Moat, H. E. S. (2013). Quantifying trading behaviour in financial markets using google trends. *Scientific Reports, Vol. 3*.

Vladimir Pestov, . (2013). Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers Mathematics with Applications*.

Vu Tien-Thanh, Chang Shu; Ha Quang Thuy, C. N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter. *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*.

Wei Fan, A. B. (2013). Mining big data: Current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*.

A. Digital Appendices

- **Scraper:** A R-project that is created and used to perform the scraping of the twitter search API.
- **Forecasting:** A R-project which contains scripts that process the data and the used machine learning methods.
- **Data:** A csv-file that contains the Dow Jones intraday prices and RData files that contain the scraped tweets.