FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Vorhersage von Wirtschaftsindikator mittels Sentimentenanalyse von Nachrichtenartikeln und maschinellem Lernen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Christoph Hämmerle, BSc
Matrikelnummer 01226577

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Allan Hanbury
Mitwirkung: Dr. Navid Rekabsaz

Wien, 5. März 2019

Christoph Hämmerle	Allan Hanbury

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Prediction of an Economic Indicator using Machine Learning and Sentiment Analysis of News Articles

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Christoph Hämmerle, BSc
Registration Number 01226577

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Allan Hanbury
Assistance: Dr. Navid Rekabsaz

Vienna, 5th March, 2019

Christoph Hämmerle   Allan Hanbury

Erklärung zur Verfassung der Arbeit

Christoph Hämmerle, BSc
Weyringergasse 31/8, 1040 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 5. März 2019

Christoph Hämmerle

Danksagung

Zuerst möchte ich mich bei meiner Familie bedanken. Sie machte es mir überhaupt möglich das Studium zu absolvieren, nicht nur aus finanzieller Sicht sonder auch die ständige Unterstützung bei auftretenden Problemen. Weiters möchte ich mich bei allen bedanken die mich während dieser Zeit begleitet und vor allem unterstützt haben, sei es fachlich oder emotional.
Ganz besonders will ich mich auch bei meinen Betreuern Allan und Navid bedanken. Mit Ihrer Hilfe wurde diese Arbeit zu dem was sie jetzt ist. Durch zahlreiche Rückmeldungen und Vorschläge die Arbeit zu verbessern und auf eine höhere Stufe zu heben, habe ich sehr viel gelernt.

Christoph

Acknowledgements

First, I want to thank my family. Without them my studies would not have been possible. They not only supported me financially but also emotionally. Their constant assistance made this possible. Further, I would like to thank everyone who accompanied and supported me on this path.
Allan and Navid, I would like to thank you for the constant help during this thesis. The countless feedback and advice made this thesis to what it is now. Through your support I was able to learn a lot.

Christoph

Kurzfassung

Das Ziel dieser Diplomarbeit ist es den Zusammenhang zwischen Meinungen und Stimmungen in Zeitungsartikeln und dem Kursverlauf des Austrian Traded Index (ATX) zu untersuchen. Dafür wenden wir bewährte Methoden aus den jeweiligen Forschungsgebieten an. Mithilfe eines Webscrapers extrahieren wir Zeitungsartikel von der Website einer österreichischen Tageszeitung. Die Polarität und die Stimmungsbilder in diesen Artikeln werden automatisiert untersucht und ermittelt. Dafür erstellen wir ein Lexikon, welches positiv und negativ assoziierte Wörter im Hinblick auf die Wirtschaft und die wirtschaftliche Entwicklung enthält. Relevante Wörter aus einem bekannten deutschen Lexikon, das SentiWS Lexikon, werden von uns extrahiert und zusätzlich durch Wörter, welche als positiv bzw. negativ im Bezug auf wirtschaftliche und finanzielle Entwicklung interpretiert werden, ergänzt. Wir erstellen weiters ein Model um Preisschwankungen und Volatilität des ATX vorherzusagen. Die Arbeit bestätigt, dass die Zeitungsartikel nützliche Informationen, um den Kursverlauf des ATX zu erklären, enthalten. Das erstellte Lexikon ist nicht eindeutig besser als das generelle SentiWS Lexikon. In einigen Testszenarien hat es jedoch durchaus Vorteile. Weiters bestätigt diese Arbeit die Ergebnisse anderer Arbeiten auf diesem Gebiet. Vor allem negative Aussagen werden erkannt und werden als einflussreich in den verwendeten Modellen erachtet. Im Vergleich dazu können positive Aussagen kaum mit positiver Entwicklung des Kurses in Verbindung gebracht werden. Die Resultate der getesteten Modelle sind stark durch diese Negativität beeinflusst.

Abstract

The analysis of textual data and their predictive quality has gained the interest of many researchers, especially in the financial domain. This thesis investigates whether newspaper articles contain information to describe the changes of the Austrian Traded Index (ATX). We apply state of the art methods to extract newspaper articles from the online platform of an Austrian newspaper, to perform sentiment analysis of the articles and to build machine learning models in order to predict price and volatility developments of the ATX. As the newspaper articles are written in German, we create a new sentiment lexicon, called German Financial Sentiment Lexicon (GFSL), by extracting sentiments from the SentiWS, a general German sentiment lexicon, and adding financial sentiment words to the lexicon. Our findings show the newspaper articles contain information which allow predictions of price and volatility movements. The GFSL does not clearly outperform the SentiWS lexicon, although in some scenarios it clearly has an advantage over the general lexicon. The results confirm the findings of previous studies such that negative sentiments highly influence the outcome of the model while positive sentiments are hardly relatable to positive development of the index.

Contents

CHAPTER 1

Introduction

Economic development is considerably influenced by sentiments, opinions, expectations and events happening around the world. The behavior and actions of individuals may be related to some force or effect which influenced them. Often newspapers are the source of information in the daily lives of people. Some articles communicate positive feelings and other negative feelings to the reader. Of course different people are not influenced the same by one article and some articles might not cause any reaction in an individual. However, the general mood, communicated by the article, can be captured via sentiment analysis. In particular, the behavior of participants in the stock market may be influenced by newspaper articles and the aggregated effects they take may be correlated to the sentiments communicated via the newspaper articles. In this case, the aggregated effects reflect the development of the economic indicators, and especially in this thesis the Austrian Traded Index (ATX).

The ATX, like every stock market index, is a general indicator of growth or decrease of the economy it represents and reflects the general mood of stock market participants regarding future developments. This thesis investigates the extent to which the communicated sentiments of newspaper articles are relatable to the development of the ATX.

Section 1.1 describes the research questions of this thesis. In Section 1.2 we present the contributions of this thesis, followed by a brief description of the methods we use. In the last part of this chapter, Section 1.3, we provide an overview of the thesis.

1.1 Research Questions

The analysis of textual data and their predictive quality has gained the interest of many researchers, especially in the financial domain. The research in this area is mostly based on the analysis of texts written in English. However, there is a lack of research on German text in the financial domain.

The first research question investigates whether the extracted sentiments correlate with

the ATX observations. We test if the sentiment weights add relevant information to the machine learning models in order to explain the development of the ATX.
For the second research question we investigate which kind of data, extracted form the ATX observations, is best described by the sentiment of the newspaper articles. We consider volatility and price movements. We also classify these time series into two groups, up and down movements, and investigate the performance of classifying price and volatility movements by training supervised machine learning models with the sentiment weights of the newspaper articles. With the models we predict the time series and compare the predictions to the actual observations of the index. We answer which kind of data, extracted from the price observation of the ATX, is best described by the sentiment weights of the newspaper articles. Do the sentiments of the newspaper articles allow qualitative predictions?
For our third research question we compare the performance of the two lexicons according to the accuracy of their predictions, resulting from the supervised machine learning models. Does the domain specific lexicon provide advantages over the general sentiment lexicon? Previous studies show the importance of domain-specific lexicons in sentiment analysis of financial data. We assume the domain specific lexicon has an advantage over the general lexicon as it provides domain specific expressions and therefore is enabled to capture the communicated sentiments more accurately resulting in higher predictive performances.

1.2 Contributions

This thesis contributes to the research area of sentiment analysis of German financial texts in two ways.
First, the creation of the German Financial Sentiment Lexicon (GFSL), a sentiment lexicon which contains words that are associated to a sentiment in the financial domain. Creating a lexicon in the German language, considering the financial context, is an important issue. The quality of the predictions from supervised machine learning models highly depends on the lexicon used to obtain the sentiment weights as the lexicon is the essential component to capture expressions in the newspaper articles. We create the GFSL by extracting words, associated with the financial domain, from the SentiWS lexicon, a well established general German sentiment lexicon [RQH10]. Further, we extend the GFSL with expressions from General Inquirer dictionaries by manually translating economic and financial expressions. We also manually add sentiment words from articles published on the Vienna stock exchange website. The GFSL lexicon is evaluated by comparing the predictions, obtained from the machine learning algorithms, to the predictions computed using the general SentiWS lexicon for sentiment analysis.
In the second contribution, we explore the applicability of newspaper articles written in German as a source of information to explain the changes of the Austrian Traded Index (ATX). The outcomes contain results of several machine learning algorithms applied to data extracted from the ATX observations, considering price differences and volatility. We investigate whether the sentiments communicated in the newspaper articles allow

predictions of the stock market price and the volatility by comparing various machine learning algorithms according to the accuracy of their predictions. We apply state of the art approaches to our problem of extracting newspaper articles, performing sentiment analysis and building machine learning models. Each step is explained in the following enumeration.

1. **Data Accumulation**
   To obtain relevant newspaper articles we implement a web scraper. The web scraper extracts data from websites, in this case newspaper articles from the online archive of *derStandard*, an Austrian daily newspaper. The need for searching online archives and solely obtaining articles which may have an impact on market development, especially articles categorized as economic, requires the implementation of a web scraper. Further, we process the extracted data and store it in a database in order to continue with the sentiment analysis.

2. **Sentiment Analysis**
   We use sentiment analysis to extract the attitude provided by the newspaper articles gathered in the first step. In order to do so we compare the words of an article to the GFSL and the SentiWS lexicon, a general German sentiment lexicon. Further, we process every newspaper article, count the occurrences of the words in the lexicons in the newspaper articles and weight these occurrences according to common weighting schemes. We use the resulting time series of weights in the third step.

3. **Machine Learning**
   Finally, we select supervised machine learning algorithms to examine the relationship between the obtained sentiment weights and the Austrian Traded Index (ATX). We investigate various settings to find the data which promises the most accurate results and allows to draw conclusions about the relationship between the sentiment weights and the ATX. We test classification and regression scenarios like price changes, volatility changes and volatility directions. The quality of the classification models is evaluated by comparing the results to the most frequent group classifier. The regression cases are tested regarding their significance. The quality of the predictions is evaluated using cross-validation techniques.

The aim of this thesis is not to provide an exact estimation of the ATX with the aid of newspaper articles and the sentiments extracted from them since the index is influenced by various factors. The work should be rather seen as an attempt to verify whether this approach leads to insights in the extraction of daily newspaper articles, written in German, from the online archive of an Austrian newspaper and whether their sentiments are relatable to economic development.

1.3 Structure of the Thesis

In Chapter 2, we describe the **Background** for this thesis. We explain the concepts of machine learning and sentiment analysis. The various approaches are briefly discussed and illustrated with examples. The examples also provide an insight into the wide area of applications.

In Chapter 3, we explain **Related Work**. Other research and its most important results are presented.

In Chapter 4, we explain the creation of the **German Financial Sentiment Lexicon** used for the sentiment analysis of the newspaper articles. We explain the process of creating the GFSL and our motivation behind it.

Chapter 5 explains the details of our approach to **Sentiment Analysis of the Newspaper Articles**. First, we explain how we accumulated the data. We then introduce our approach to perform the sentiment analysis of the newspaper articles. In the last section of this chapter we describe the design of our experiments.

In the first part of Chapter 6, **Experiment Results**, we explain and illustrate the outcomes of the experiments. In the second part we qualitatively analyze the results of the experiments. Finally, we discuss the main findings.

The last chapter, **Conclusion and Future Work**, summarizes the outcomes and conclusions of this thesis and presents further research possibilities in this area.

CHAPTER 2

Background

Chapter 2 provides the background needed for the analysis of the newspaper articles. Section 2.1 provides an introduction to machine learning. Different types of machine learning and their applications will be discussed. Furthermore, the general approach to build and evaluate a machine learning model is briefly discussed. In Section 2.2 we explain sentiment analysis. We define the term sentiment analysis and introduce various areas of application. Finally, we introduce different approaches and levels of analysis to illustrate common challenges in this research area.

2.1 Machine Learning

Machine learning is a research field which addresses extracting knowledge from data. Machine learning algorithms aim to learn patterns in the provided input data in order to generalize decision making processes to unseen data. Manually designing rule-based systems is feasible for applications in which humans have a good understanding about the rules and the general functionality which they aim to describe. Nevertheless, there are disadvantages regarding rule-based systems. For example the defined rules are domain specific. Changing the domain slightly might result in a complete rework of the defined decision rules. Further, defining rules requires a deep understanding about the domain. Machine learning provides the tools to automate the decision-making process by generalizing from input data [MG16].

In the first part of this section we describe the different types of machine learning. Examples illustrate their area of application. The second part of this section describes the general machine learning process to build a machine learning model. Finally, we briefly explain the machine learning algorithms used for this thesis.

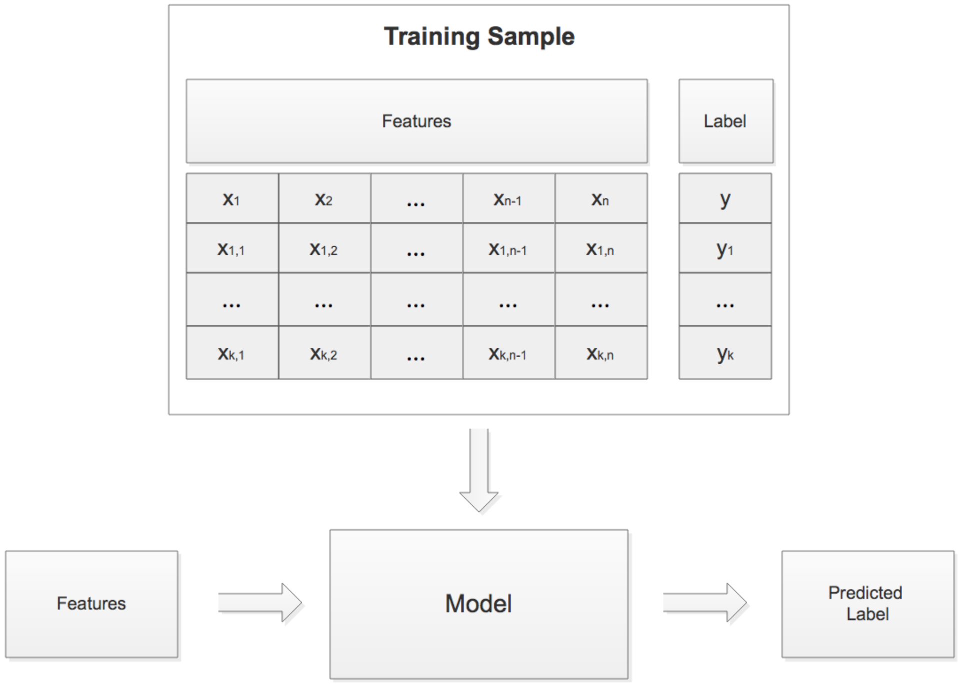

Figure 2.1: Supervised Learning

2.1.1 Types of Machine Learning

Raschka [Ras15] distinguishes between three types of machine learning: supervised learning, unsupervised learning and reinforcement learning. The fundamentals and the differences between these types are explained in this section. Examples will illustrate the application area of the different types of machine learning.

Supervised Learning

The goal in supervised learning is to learn a model with given data and given labels. The features describe the given input data and the label represents the output of the model. This means the model is trained with input values and the corresponding given output value. The trained model is then used to process unseen data and predict the output. The term supervised refers to a set of samples where the desired output is already known and used to train the model [Ras15]. Figure 2.1 visualizes the approach. The first row describes the features names $X_1, ..., X_n$ and the label name y. The consecutive rows display the features of different observations. The model is trained with the feature set and the corresponding label set. Each observation consists of features and the corresponding output. After the model is trained the same features are extracted from

unseen input data. The supervised machine learning model is then enabled to predict the corresponding label.
As explained by Raschka [Ras15], the supervised machine learning methodology consists of two major subcategories. The first is classification. The goal is to predict categorical class labels of new data based on past data. An example for multi-class classification would be the recognition of letters from handwriting. The task is to classify handwritten letters into letters from the alphabet. To do so an algorithm is trained with several examples of handwritten letters and the corresponding correct classification. The algorithm should then be able to classify new observations with a certain accuracy. An example of a binary classification problem is the identification of spam emails. An algorithm is trained with emails and the corresponding categorization, either is spam or is not spam. According to the defined features the algorithm should then be able to classify future observations.
The second type of supervised machine learning is regression. The goal of regression analysis is to predict a continuous outcome from a given number of explanatory variables by finding a relationship between the explanatory variables and the response variable. An example for a regression model is the prediction of someones annual income with the level of education, the age and the place of living. The procedure to do so is the same, except for the algorithm, as for classification. The algorithm is trained with the feature data and the corresponding label, in this case education, age and place of living and the corresponding annual income. The algorithm is then able to predict from future observations [MG16].
In this work, we apply both supervised classification and regression machine learning methods to investigate the relationship between the newspaper articles and the changes of the ATX.

Reinforcement Learning

Reinforcement learning deals with the continuous improvement of a system based on the interactions of the system with the environment. The system has a current state and a reward signal. Through interaction with the environment the system learns a series of actions, via exploratory trial-and-error approach or planing, in order to maximize the reward signal. An example for such a system would be a chess player simulation. The system learns to perform moves according to the current state and the likely reward, in this case to win the game [Ras15].

Unsupervised Learning

Unsupervised learning is used to explore the structure of the data and to extract meaningful information without training the algorithm with a known outcome or defining a reward signal. Clustering and dimensionality reduction are two major applications.
Clustering is a technique to organize observations into groups which share a degree of similarity and are dissimilar to other groups. This technique is often used for marketing purposes to group users and target the different groups individually [Ras15].
Dimensionality reduction is another important application of unsupervised learning. Data

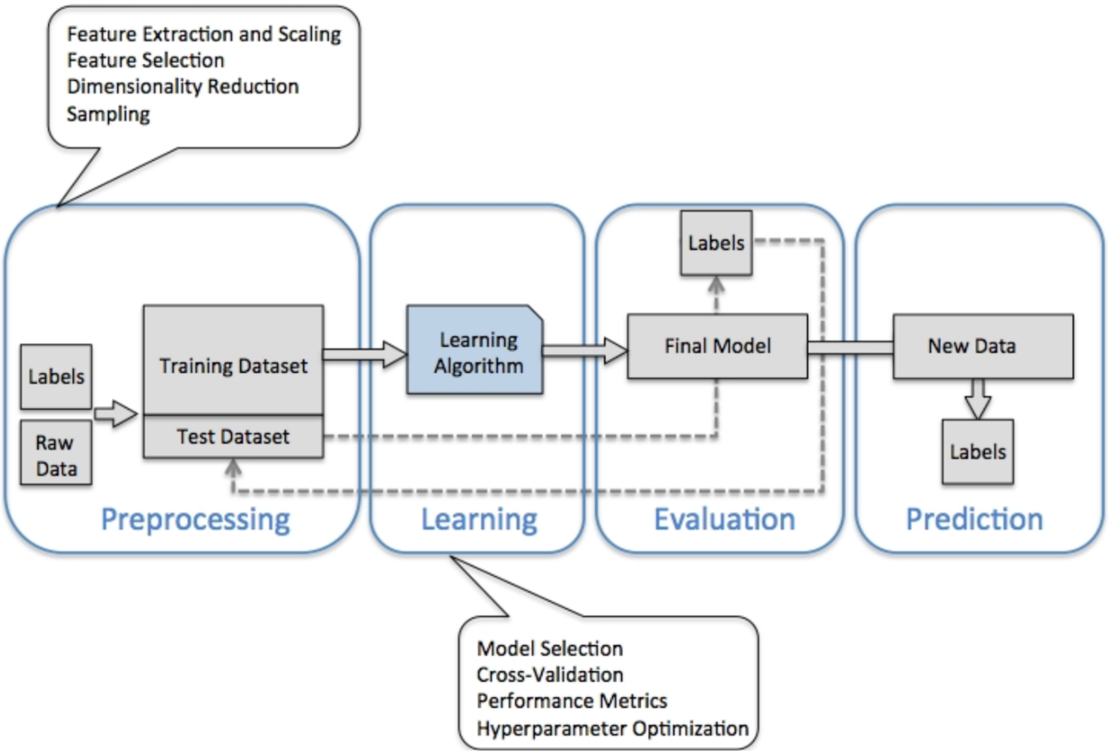

Figure 2.2: Machine Learning Process [Ras15]

often consist of multiple features. Due to limitation of computational power, storage capacity, computational time or to reduce noise in data the reduction of dimensions is helpful. The projection of multiple dimensions into two or three dimensions is useful to visualize data [Ras15].

2.1.2 Machine Learning Process

This section is about the process to train and evaluate a machine learning model in order to be able to accurately predict new data. We summarize the machine learning process defined by Raschka [Ras15], Figure 2.2, in the following.

Preprocessing

The preprocessing is about preparing the input data. As described in the supervised machine learning approach the data consist of the feature set and the label set. The feature set is used to describe and explain the behavior of the labels. The label set is the data which is intended to be described by the model. Feature extraction deals with the creation of features based on the input data. The features may differ regarding their value range and or distribution. In order to prevent unequal influence of features, based on their value range, on the machine learning algorithm, especially algorithms which

compute distances between features, the features might need to be scaled.
Features might not always contain useful information or the information of a feature is described by other features. In such a case it is desired to only use the features which indeed are useful to explain the label. Further, the feature set is often a high dimensional structure. Features often correlate with each other or do not provide information in order to explain the corresponding label. Therefore, it is important to reduce the dimension of the features. This reduction results in a less complex structure and allows future steps, like training the machine learning algorithm, to use less computational power and time. The reduced amount of dimensions may be varied to see whether a higher number of dimensions will increase the accuracy significantly or if the chosen reduction justifies the possible loss in accuracy.
In order to train and evaluate the machine learning model, the available data is split into two sets, called training- and testing set. The training set is used to train the algorithm and the testing set is used to evaluate the quality of the model. In order to test different settings a part of the training set may be further split. The resulting set is referred to as the validation set. The validation set is used to optimize hyper-parameters. Once the most promising setting of hyper-parameters is found, the model is applied on the testing data. The result of the testing data is the indicator of the quality of the model.

Learning and evaluation

The learning process of the algorithm highly depends on the previously processed data. Different feature sets and labels may lead to many input combinations for a machine learning model. Testing various feature sets with the corresponding label, which together with a suitable algorithm, provide accurate predictions is challenging and may take several attempts.
In order to compare the performance of machine learning models we use common metrics. For the regression algorithms we use the R-squared metrics to compare different experiments. The R-squared value provides information about how close the predictions compared to the actual observations are. It is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\hat{y_i}$ are the predicted values and y_i the actual observations. The numerator is referred to as the residual sum of squares, it describes the difference between the actual observations and the predictions. The value of the R_2 is between 0 and 1. If the residual sum of squares is zero, the predictions are equal to the observations, the fraction will be zero and the R_2 has its maximum value of 1. The denominator describes the variance of the actual observations.
In order to compare different classification approaches we use the prediction accuracy which is defined as the number of correct classifications divided by the total number of classifications.
Machine learning algorithms often consist of many hyper-parameters which affect the

behavior of the machine learning algorithm. These hyper-parameters have to be optimized to the given data. In order to optimize the hyper-parameters to our data we use grid search with cross-validation. Grid search is a technique to test all combinations of predefined parameter settings. For each parameter we define the settings we want to investigate. The machine learning algorithm is trained and evaluated with every combination of the parameter settings resulting in the optimal parameter values. It is crucial to avoid overfitting the model. This means to overly optimize the parameters to the specific input set. If the results of the validation and the testing set are highly different the model may be overfitted, meaning the hyper-parameters are tuned too much. Applying the algorithm on new data may then result in poor predictions although the algorithm provided good results on the validation set. Testing the final model on the test set indicates whether the model is able to adapt to new data. Raschka [Ras15] mentions in this context the term *generalization performance*. This means the algorithm has to have the ability to not only fit the training data but also to process new data and receive comparable results. If the performance is not satisfying or the model is overfitted the procedure starts from the beginning.
We retrain the resulting model with the optimized hyper-parameters on the training and validation set to increase the amount of training data. We compare the predictions to the test set of the label and compute the performance metrics. Furthermore, we provide information about the significance of the used model, answering if the features add useful information to the model. For the regression case the information about the significance of the results may be obtained via the p-value with a defined significance level. The classification results are evaluated using the different splits of the input set. The results are compared to a baseline. A baseline is a trivial classifier, for example a classifier which always predicts the most common group in the training set.

Prediction

After a suitable model is found and the performance is satisfying the model may be applied on new future data.

2.1.3 Algorithms

In this section we briefly explain the algorithms applied in this work. We provide sources which explain the algorithms in detail.

- **Support Vector Regression (SVR):** In the following we will briefly explain the basic functionality of SVR with the example in Figure 2.3. In this example the problem is reduced to a two dimensional linear regression problem. The goal is to find a function $f(X)$ that has the highest ϵ value, including many observations, and at the same time is as flat as possible. So observations with an error less then ϵ, between the tube $+\epsilon$ and $-\epsilon$ and the function $f(X)$, will be tollerated. Observations outside the tube, with a higher error than ϵ, will be considered in the optimization function. *"The constant $C > 0$ determines the trade-off between the*

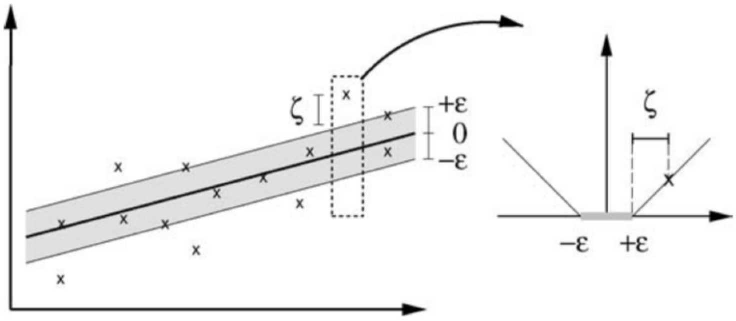

Figure 2.3: Linear SVR, Smola and Schölkopf [SS04]

flatness of f and the amount up to which deviations larger than ϵ are tolerated." So C can be seen as the penalty parameter for the error term. The observations outside the tube are penalized and contribute to a cost function which is minimized in the optimization problem [SS04].
This example is very simple but sufficient to explain the basic functionality. The data is not always linear, as in this example. To cover non-linear data the kernel can be varied. The kernel is used to map the input data into higher dimensional spaces in which the data is separable. The line may then be a curve but the general idea is the same. Further, the example is only two dimensional. Increasing the dimensions increases the complexity of the optimization problem. The function $f(X)$ then describes a high dimensional construct instead of a line. For a detailed description of SVR we refer to Smola and Schölkopf [SS04].
The SVR algorithm has various parameters to control the learning process. First, the kernel types like linear, polynomial, sigmoid and rbf. The penalty parameter, C, the parameter defining the size of the tube, ϵ, in which no penalty is assigned and the parameter regulating the tolerance, *tol*, for the algorithm to stop.[1]

- **Support Vector Classification (SVC):** SVC is very similar to the previously described SVR algorithm. As mentioned by Hsu et al. [HCL03] the goal is to find a linear separating hyperplane with the maximal margin. Figure 2.4 illustrates the desired outcome in a two dimensional example. The different shapes of the observations signalize the group affiliation. The goal is to find a hyperplane which separates the groups, in this example the dashed line with the additional maximal margin. Similar to the SVR algorithm different hyper-parameter settings can be selected, like the kernel and again the penalty parameter C.

- **Random Forest Regression (RFR):** Breiman [Bre01] provides a detailed description of the Random Forest algorithm. We will briefly explain the functionality

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html`, last accessed 05.09.2018

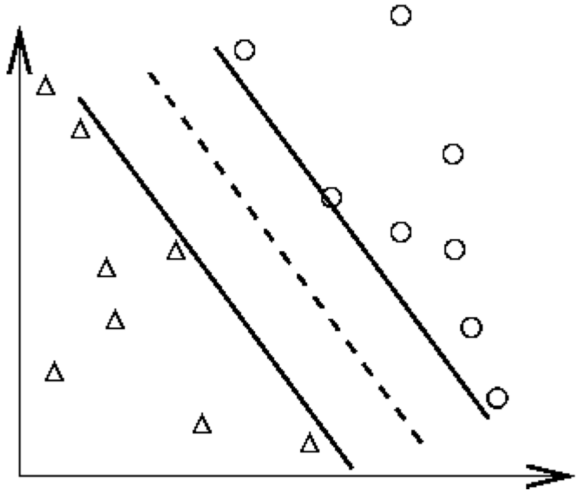

Figure 2.4: SVC Example [HCL03]

according to the documentation of the used implementation.[2] The random forest algorithm follows an ensemble approach. The goal is to combine the predictions of several base estimators, in this case decision tree estimators, to improve generalizability and robustness of the estimation. Each tree in the ensemble is built from a sample drawn with replacement from the training set. Randomness is added to the algorithm by choosing the best split among a random subset of the features. In the regression case the final estimation is calculated by averaging the output of all trees in the ensemble.

- **Random Forest Classification (RFC):** The base functionality of the algorithm is similar to the regression case. The final decision to which class the observation is assigned to is conducted via averaging the probabilistic prediction of all classifiers.[2]

- **Principal Component Analysis (PCA):** Jolliffe [Jol11] describes PCA as probably the most common technique used for dimensionality reduction. PCA reduces dimensionality by finding linear combinations, called principal components, which have maximum variance for the data. Furthermore, the components are uncorrelated in order to minimize loss in information and maximize the dimensionality reduction. Jolliffe [Jol11] provides a detailed mathematical explanation.

2.2 Sentiment Analysis

Liu [Liu15] defines sentiment analysis as follows: *"Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals,*

[2] `http://scikit-learn.org/stable/modules/ensemble.html#forest`, last accessed 23.10.2018

attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics. [...] In a nutshell, sentiment analysis or opinion mining aims to identify positive and negative opinions or sentiments expressed or implied in text and also the targets of these opinions or sentiments [...]."

Since the primarily focus of sentiment analysis is on analyzing textual data it has been a research area of natural language processing (NLP). Sentiment analysis has also been widely studied in the context of data mining, web mining, and information retrieval. Due to the increasing importance of sentiment analysis to businesses and society, the research area has spread from computer science to management and social science [Liu15].

In the following we provide an overview of the broad research area of sentiment analysis and opinion mining. First, the development and some applications are presented. We further explain common approaches for sentiment analysis, different levels of analysis and the processing of text prior to sentiment analysis. At the end of this section we highlight common challenges in sentiment analysis to show the complexity of this research area.

2.2.1 Development and Applications

The increasing amount of data and more importantly the availability of data in digital form led to an increase in research on sentiment analysis. Opinions of individuals are widely spread among blogs, reviews, social media, comments, and every form of participation or engagement in a topic. Sentiment analysis not only focuses on the attitude which a text transmits but also on the individual who wrote this text. Based on the statements in the text a sentiment profile of the author may be established. This information provides various possibilities like product recommendations or the determination of the political orientation of the individual.

Not only businesses and organizations are interested in the opinion about their products and services but also governments have interest in how the public thinks about policies and policy changes. It is no longer necessary to conduct a survey about how users think about a specific topic. This information may already be publicly available on social media or other platforms. Further, there are various research topics which use sentiment analysis. For example during elections, the posts on Twitter provide information about candidates approval. Another popular application is stock market prediction with the aid of public opinions, CEO - letters, board posts, and other publicly available information concerning this domain [Liu15].

2.2.2 Common Sentiment Analysis Approaches

This section is dedicated to introduce common approaches to perform sentiment analysis. In the following we briefly summarize the approaches listed by Cambria et al. [CDBF17]:

- **Knowledge-based technique:** Text is classified according to unambiguous affect words. Basically, a predefined set of words is matched to a text, trying to determine its sentiment. The major weakness of the knowledge-based approach is poor recognition when linguistic rules are involved. For example, the sentence *"Today was a happy day"* can be correctly classified by a knowledge base while *"Today wasn't a happy day at all"* will likely be incorrectly classified. The accuracy of knowledge-based approaches highly depends on the employed resources. Without a comprehensive knowledge base it is difficult to determine the sentiments associated with natural language.

- **Statistical methods:** Statistical methods intend to learn grammatical constructs and word co-occurrence frequencies. Techniques like deep learning use large training corpora of affectively annotated text to learn the valence of affect keywords and word co-occurrence frequencies. Due to the amount of data statistical methods need they are only accurate when applied to users'text on the page or paragraph level.
- **Hybrid approaches:** Hybrid approaches exploit both knowledge-based techniques and statistical methods to better grasp conceptual rules and language constructs to accurately classify text. These approaches try to combine the advantages of recognizing semantics of knowledge-based approaches and the ability to recognize linguistic patterns of statistical methods in order to better understand the sentiments transmitted.

Bing Liu [Liu15] distinguishes between unsupervised and supervised methods for sentiment analysis. The summarized descriptions of both approaches are listed below.

- **Supervised Classification:** Sentiment classification is a text classification problem. Researchers applied machine learning algorithms to classify text. The key for accurate classification are the features. The most common used features are terms and their frequency. These features are individual words and their frequency counted in the text. Furthermore, the part of speech (POS) is considered as another class of feature. For example, adjectives are important expressions of opinions. Phrases which are used in natural language are challenging to automatically capture. There are also words which completely shift the sentiment of a text, the classical example is *not*, or *don't*. Several studies use different combinations of these features for their machine learning algorithms.
- **Unsupervised Classification**: The unsupervised approach uses phrases and sentiment words to classify text. Syntactic patterns are analyzed in the text and interpreted. Therefore, a sentiment lexicon containing words and phrases which are relatable to the investigated sentiments is used. A sentiment score is then computed for each document according to the expressions found in the text and the scores which were assigned to each expression. The simplest form to compute the score for a document is to add up all scores for the expression and check whether this score is positive or negative. Of course, there are many variations to this approach, not only in calculation of the overall score but also the score of each expression in the text.

2.2.3 Levels of Analysis

In this section we highlight the different textual levels of analysis. The different levels are described by [Liu15].

Document level

The document level tries to classify whether a document as a whole expresses a positive or negative sentiment. This analysis implies that each document describes the opinion of one entity, for example a product review describing the opinion about one product. Further, there exists one opinion holder, for example one person which wrote the product review. These assumptions are necessary, multiple opinion holders in one document may have different opinions about the underlying entity it describes. This example already describes a limitation of document analysis, for example if a text contains sentiments describing multiple entities a more fine-grained analysis is needed. The foundation for this kind of analysis is described in the sentence level and the aspect level.

Sentence level

The previous example described the need of a more detailed analysis of a document. This level analyzes each sentence in a document. The general goal is the same, classifying sentences into sentiments. Each sentence can be seen as a very short document. The challenge is the reduced amount of information. Therefore, sentences may also contain no specific sentiment at all and need to be seen as neutral, for example facts. Furthermore, the difference between subjectivity and objectivity has to be considered in the analysis. Subjective sentences express sentiments of an opinion holder whereas objective sentences are expressed by someone but display a general statement, this statement although can have some implicit sentiments. For example, the sentence *"The earphone broke in two days."* expresses a fact but implicitly describes an unsatisfying condition.
The sentence level analysis assumes there is one sentiment per sentence. The target of the sentiment is not considered, which displays some restrictions in this approach. For example, the sentence *"The picture quality of this camera is amazing and so is the battery life, but the viewfinder is a little small for such a great camera"* expresses positive sentiments for the picture quality and the battery life whereas the viewfinder is viewed as too small. Sentences may contain different sentiments for different targets. The aspect based level tries to correctly classify opinions in a text and further determine also the target of these sentiments.

Aspect level

The aspect level further tries to extract the target of the sentiment expressed by the holder of the opinion. Or to whom/what does the expressed sentiment belong? For example, it does not make sense to declare the sentence *"Apple is doing very well in this poor economy."* as either positive or negative. This sentence contains two contrary sentiments. The aspect level focuses on extracting the sentiment and matching it to the right target. In this example *doing very well* to *Apple* and *poor* to *economy.* This example shows the complexity of sentiment analysis.

2.2.4 Text Processing

Preprocessing text is an essential task prior to performing sentiment analysis. Pröllochs et al. [PFN15] list the following steps:

1. **Cleaning:** Punctuation marks and other symbols are removed from text.
2. **Tokenization:** Text is mostly stored as a string. The analysis of the texts is simpler if the text is stored as single words. Each sentence is split into it's consistent words. These single words are referred to as tokens.
3. **Removal of stop words:** Stop words are words without a deeper meaning, such as the three articles of German *der*, *die*, *das.* Due to their missing sentiment and irrelevance in the context of sentiment analysis they can be omitted.
4. **Stemming:** Stemming describes the process of reducing a inflected word to it's stem. The goal is to capture the meaning of inflected words even if the stem is not a valid root form.

2.2.5 Challenges in Sentiment Analysis

Language nuances, like language-specific phrases and the use of grammatical constructs, which may seem easy in daily use, intensify the complexity an algorithm faces to perform sentiment analysis. This section briefly describes further challenges in analyzing language constructs.
Negations and even double negations completely change the sentiment of a text. For example the sentence *"Das Produkt ist nicht unbrauchbar."* uses the word *nicht* to shift the sentiment, initially negative, of the word *unbrauchbar* to an overall positiv sentiment of the sentence.
Further, sarcasm and ironic expressions, which can be easily understood by humans, are a complex task for an algorithm to identify. Imagine a product review of a customer. The customer describes the incompetence of the service team and writes the following sentence: *"Bei der hohen Qualität des Kundenservices muss das Produkt mindestens genau so gut sein."* Humans easily understand the customer claims the product has to be equally bad as the service provided by the company. For an algorithm this is not trivial to understand.
There are many texts which require sensitivity to capture essential and important parts. Some expressions are not as relevant as others which reveal themselves only by reading and understanding the full text.
Combining all the mentioned language-specific phrases in a single text requires the algorithm to not only correctly identify these constructs but also to interpret them accordingly in order to accurately classify expressions.

CHAPTER 3

Related Work

Chapter 3 provides an overview about related work and contributions of researchers in the area of sentiment analysis, focusing on the financial domain. The first section, Section 3.1, summarizes the outcomes of other research in this area. In Section 3.2 we list German resources to perform sentiment analysis.

3.1 Sentiment Analysis in the Financial Domain

Understanding the market dynamics, which tend to explain the behavior of markets, has gained significant interest in research and industry. Newspapers and other forms of media transmit information to consumers. Consumers are often not aware of how these underlying sentiments of news may affect their behavior. Tetlock [Tet07] investigates dependencies between media and stock market activities. Negative sentiments of media induce downward pressure on prices. Furthermore, pessimism leads to temporarily high trading volume. The temporary decrease in returns, which can be predicted by pessimism, are reversed within a few days.
Schumaker and Chen [SC09b] distinguish two different trading philosophies, the fundamentalists and the technicians. The first group tries to determine the price of a stock from financial numbers of the overall economy or the specific sector in which the stock is traded. Contrary supporters of the technician philosophy try to derive the price from historical data. They believe arbitrage opportunities can be found by sensitively analyzing the historical data and volume movements. Furthermore, Schumaker and Chen compare different approaches to predict prices from financial news articles. They claim the most common technique used to classify articles is the bag of words approach. This technique simply searches for the occurrences of words in the article and assigns a sentiment to it. In their approach, grammar or specific word combinations, which combined establish a different meaning, are ignored.
The importance of grammar and phrases are addressed by Chan and Chong [CC17].

They recognize phrases in news articles by a machine learning technique. They conclude the sentiments expressed through text streams are helpful to analyze trends in a stock market index.

Schumaker et al. [SZHC12] try to answer if positive and negative subjectivity influence the prediction of stock prices with news articles. Especially negative subjectivity of news articles have an impact on trading behavior. This conclusion supports the findings of Akhtar et al. [AFOS13]: *"We document asymmetric announcement effects of consumer sentiment news on United States stock and stock futures markets. While a negative market effect occurs upon the release of bad sentiment news, there is no market reaction for the counterpart good news."*

Nopp and Hanbury [NH15] analyzed risk attitudes of banks with sentiment analysis of CEO letters and outlook sections of annual reports. They recommend sentiment analysis for detecting risks to be used in an aggregated form since the evaluation on the individual level led to inaccurate predictions contrary to the aggregated analysis which revealed significant correlations.

While the previous studies focus on the sentiment, extracted from text, Rekabsaz et al. [RLB+17] explores the effect of combining factual market data and sentiment scores. They evaluate different fusion techniques to combine factual data and text resources. Further, they predict the volatility in financial markets with factual market data and sentiment scores of annual disclosures of companies.

Several studies address the possibilities of stock market prediction using textual data. Henrique et al. [HSK18] use Support Vector Regression (SVR) to predict stock prices for small and large capitalizations by using daily and up-to-the-minute price information. They claim the model has predictive power especially when using a strategy to periodically update the model. Their findings also indicate an increased accuracy of their predictions during lower volatility periods.

Atkins et al. [ANG18] used two stock indices and two equities to empirically show that information extracted from textual news sources is better at predicting the volatility direction of the market than the price movement. *"Though the inability to predict close price movement any better than random contradicts previously published results, our results suggest that information in news, influencing markets via sentiment-driven behavior essentially affects second order statistics of the financial system."* Regarding second order statistics of the financial system they mention volatility and trade volumes. For their analysis they extracted news articles describing the market, politics, business and world news. The extracted articles are then preprocessed. Further, the semantics of the articles are modeled using topic models. These topic models and the volatility movements were used to fit a Naive Bayes prediction model. They managed to achieve an average directional prediction accuracy for volatility for new data of 56% while the prediction of the asset close price performed no better results than random classification.

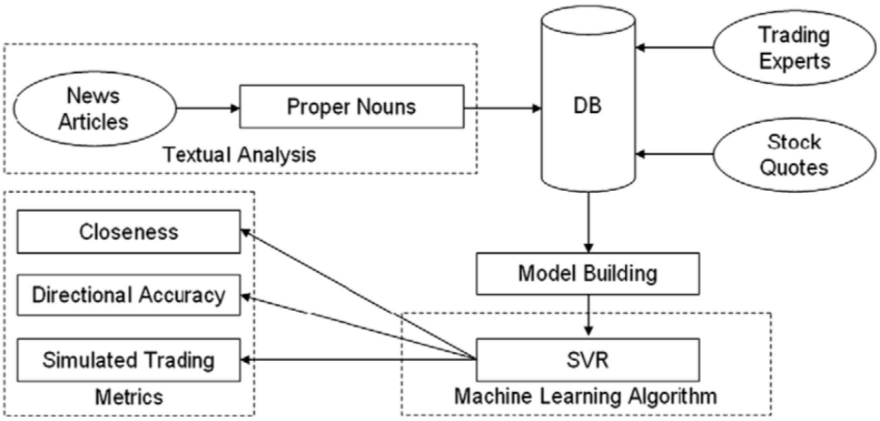

Figure 3.1: System Design of Schumaker and Chen [SC09a]

Schumaker and Chen [SC09a] provide an illustration, Figure 3.1, regarding the functionality of the approach they used to predict stock prices. They extract news articles from Yahoo! Finance and represent them using Proper Nouns. Proper Nouns retain only nouns and noun phrases, which fall in defined categories, from the news articles. Further, minutely stock price data and information regarding buy/sell recommendation from trading experts are gathered and stored. They built and tested different models according to the Global Industry Classification Standard [1] (GICS) distinguishing between sectors, industry groups, industries, and sub-industries. The models are built with only the relevant information regarding the classifications. They use the SVR algorithm with a binary representation of the proper nouns, either being present in the financial news articles or not, for their models. Training the model based on sectors resulted in a score of directional accuracy of 71.18%. Directional accuracy measured if the prediction of the price (up/down) matched the actual price movement. Further, they simulated trading and achieved 8.5% return on invest. For our approach we perform a sentiment analysis of the newspaper articles and represent each article as a vector of sentiment weights. We build a model and apply the SVR algorithms to these weights to predict price and volatility movements. As a metrics we compute the R2 value and the classification accuracy, the percentage of correctly classified observations, in order to compare the results of various models.

Luss and D'Aspremont [LD15] predict the direction of returns and abnormal returns, absolute returns greater than a predefined threshold, of intraday price movements using SVM classification. They apply a bag-of-words approach to press releases and weight the word occurrences with the TFIDF weighting scheme. Although they are unable to predict the direction of returns, abnormal returns appear to be predictable using absolute returns or textual data.

[1] `https://www.msci.com/gics`, last accessed 09.09.2018

3.2 Sentiment Lexicons

This section lists common German lexicons in the literature. Each of them is briefly described.

- **SentiWS**: The SentiWS [RQH10] lexicon is a German language resource for sentiment analysis provided by the University of Leipzig. They provide this lexicon as part of their project "Deutscher Wortschatz".[2] This project provides information about the German language since 1990. They gather and edit data from publicly accessible documents and combine them to a corpora collection. These collections are among the most comprehensive publicly available data in German.[3] The SentiWS lexicon consists of approximately 1650 positive and 1800 negative word stems. Furthermore, the lexicon contains different derivations of each word resulting in approximately 16000 positive and 18000 negative words.[4]

- **German Emotion Lexicon:** Klinger et al. [KSR16] create German emotion lexicons to automatically detect emotion in literary studies. Currently they provide lexicons for seven emotions.[5]

- **SePL:** Sentiment Phrases List (SePL) is a German list of phrases which are associated with a sentiment in German. The list is provided by the institute of information systems from the university of Hof and contains over 14000 entries.[6]

- **GermanPolarityClues:** is a German sentiment lexicon created by Waltinger [Wal10] by combining several lexicons. The lexicon has over 10000 polarity features associated to three polarity scores. The scores provide a direction, positive, negative or neutral, to the terms.[7]

- **OpenThesaurus:** A platform which provides synonyms and associations for German words. Further, they provide wordlists for specific domains like geography, biology, physics, politics, economy and more.[8]

- **BPW Dictionary:** Bannier et al. [BPW19] provide a German dictionary for business communication. The dictionary is based on a commonly used English dictionary to examine finance and accounting specific text. The BPW consists of more than 10000 words and is publicly accessible on the website of University of

[2] `http://wortschatz.informatik.uni-leipzig.de/de`, last accessed 23.07.2018

[3] `http://wortschatz.informatik.uni-leipzig.de/de/documentation`, last accessed 23.07.2018

[4] `http://wortschatz.informatik.uni-leipzig.de/de/download`, last accessed 21.03.2019

[5] `https://bitbucket.org/rklinger/german-emotion-dictionary/src`, last accessed 06.02.2019

[6] `http://www.opinion-mining.org/SePL-Sentiment-Phrase-List`, last accessed 06.02.2019

[7] `http://www.ulliwaltinger.de/sentiment/`, last accessed 06.02.2019

[8] `https://www.openthesaurus.de`, last accessed 06.02.2019

Giessen.[9] The lexicon was published after conducting our experiments therefore we do not use it in this thesis. We include it in this list to provide comprehensive sources to perform sentiment analysis for German text.

[9] https://www.uni-giessen.de/fbz/fb02/forschung/research-networks/bsfa/textual_analysis, last accessed 21.04.2019

CHAPTER 4

German Financial Sentiment Lexicon

In Chapter 4 we explain the process of creating the German Financial Sentiment Lexicon, referred to as GFSL. In Section 4.1 we explain our motivation to create the GFSL. Section 4.2 describes the sources to build the lexicon. In the last section, Section 4.3, we provide information about the creation of the GFSL.

4.1 Motivation

We believe a domain specific lexicon enables a more accurate extraction of sentiments, considering the financial domain, communicated in the newspaper articles. Domain specific expressions, which are not considered in a general sentiment lexicon, allow us to extract the sentiments more precisely. This increase in quality of the obtained information may provide the possibility to gain more accurate predictions, assuming there exists a correlation between the sentiments in the newspaper articles and the changes of the ATX. Although we found several resources to perform sentiment analysis for German, as listed in the **Related Work** chapter, we identified a lack of research in this area because there is no German sentiment lexicon for the financial domain available. Motivated by this research opportunity we have decided to create the GFSL.

4.2 Used Resources

The basis for the GFSL is taken from the SentiWS lexicon. Remus et al. [RQH10] created this lexicon. SentiWS lists positive and negative sentiment bearing words, their part of speech tag, the positive and negative polarity of the word in an interval of -1 and 1, and some of the inflections of the word. It contains adjectives, adverbs, nouns and verbs which express a positive or negative sentiment. The SentiWS lexicon is built with three major resources:

- General Inquirer: Remus et al. [RQH10] use the categories *Pos* and *Neg* and translate them via Google Translator into German. These categories provide the basis of the lexicon. Additionally, they manually added some words of the finance domain since the lexicon was initially developed for sentiment analysis of financial blog and newspaper articles.
- Co-occurrence Analysis: The second source is a co-occurrence analysis of rated product reviews. The authors identified words which occur significantly often and manually chose the 200 most significant ones.
- German Collocation Dictionary: The third source is the German Collocation Dictionary. The words extracted by the two previous sources where used to distinguish between semantic groups. The German Collocation Dictionary contains semantic groups from which additional words are extracted.

The latest version of the SentiWS lexicon consists of approximately 1650 positive and 1800 negative word stems. Including the inflections of the words, the SentiWS lexicon consists of 16000 positive and 18000 negative word forms.[1]

4.3 Creating GFSL

The initial SentiWS lexicon contains words which are uncommon in the financial and economic domain. These words are used to capture the general sentiment of text. For the GFSL we manually extract approximately 1500 basic word forms, without the inflections, related to sentiments in the financial domain by our personal judgement. Further, the polarity of the words is removed since the lexicon is adapted for a different purpose and therefore of limited use.
We extend the GFSL by context specific dictionaries from the General Inquirer Categories.[2] The General Inquirer is a computer-assisted approach for content analysis. It is basically a mapping tool which categorizes texts into domains according to predefined dictionaries which contain common words for the specific domain.[3] The words of the following categories are translated and added to the GFSL. The dictionaries can be found at the General Inquirer homepage.[2]

- Econ@: containing words of financial, economic and industrial context.
- Exch: words concerning buying, selling and trading.

[1] `http://wortschatz.informatik.uni-leipzig.de/de/download`, last accessed 21.03.2019
[2] `http://www.wjh.harvard.edu/~inquirer/homecat.htm`, last accessed 23.07.2018
[3] `http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html`, last accessed 23.07.2018

- TrnGain: words describing transaction gains, general words for accomplishment.
- TrnLoss: words describing transaction loss, general words for not accomplishing
- WltPt: words for wealth in business and commerce.
- WltTran: words for pursuit of wealth.

Further, we manually added expressions describing trends of stock market indexes, development of companies and phrases about future developments by reading financial and economic newspapers and news of the Vienna Stock Exchange.[4] We added approximately 200 words like *Kursgewinn*, *Kursverlust* and *insolvent* to the GFSL.
The GFSL contains only stemmed word forms because in German words may have many inflections. Including and finding all inflections of the words in the GFSL is a major effort. We decided to stem the words in the newspaper articles as well in order to compare the words in the GFSL to the words in the articles. To stem the words we use the German stemmer in the NLTK python library.[5] The NLTK - package uses the Snowball stemming algorithm for German language.[6] The stemmed words are further checked for uniqueness. The final GFSL contains a total of 2267 positive and negative stemmed words describing positive and negative sentiments in the financial and economic domain. The GFSL is attached at the end of the thesis.

[4] `https://www.wienerborse.at/news/`, last accessed 11.10.2018
[5] `https://www.nltk.org`, last accessed 18.08.2018
[6] `http://snowballstem.org`, last accessed 18.08.2018

CHAPTER 5

Sentiment Analysis of the Newspaper Articles

In Chapter 5 we explain our approach to perform the sentiment analysis of the newspaper articles. Section 5.1 describes the extraction of the newspaper articles. In Section 5.2 we explain the process of preprocessing and weighting the newspaper articles. Section 5.3 lists several label vectors we create to perform the machine learning. In the last part of Chapter 5, Section 5.4, we introduce our experiment design.

5.1 Data Accumulation

The news articles are gathered from *derStandard*.[1] It is an Austrian newspaper and has all of its articles available in an online archive.[2] The archive consists of articles since 2007. This online archive is the source for the data. The structure of the archive provides a possibility to navigate through the articles.

https://derstandard.at/archiv/(year)/(month)/(day)

The parenthesized expressions, e.g. year, month and day, display the structure of the URL used to navigate through the archive. By inserting the defined values a webpage containing a list of links, referring to every single article which was published at this specific date, is provided. The following section explains the algorithm used to extract the relevant articles from the archive and the challenges which occurred throughout the process.

[1] `https://derstandard.at`, last accessed 23.07.2018
[2] `https://derstandard.at/archiv`, last accessed 23.07.2018

5.1.1 Web Scraping

Web scraping is the automated process of gathering data from the internet. This is mostly accomplished by writing a program which accesses web pages and extracts the desired information [Mit15].
The following pseudocode explains the functionality of the algorithm used to retrieve the articles.

Algorithm 1 Web Scraper

```
1: startDate ← 2007/01/01
2: endDate ← 2017/12/31
3: dates ← getDates(startDate, endDate)
4: for each date in dates do
5:     html ← getHTML("https://derstandard.at/archiv/" + date.year + "/" + date.month + "/" + date.day)
6:     links ← extractAllLinks(html)
7:     for each link in links do
8:         article ← getHTML(link)
9:         if validateArticle(article) then
10:            headline ← getHeadline(article)
11:            text ← getText(article)
12:            storeArticle(date, headline, text)
13:        end if
14:    end for
15: end for
```

The web scraper accesses every date in the archive which lies in between the *startDate* and *endDate*. The function *getDates*() returns all relevant dates, including the starting date and the last date of which the articles should be extracted. The URL for every date is accessed and the HTML code is saved in *html*. This HTML code contains a list of links to the articles which were published at this date. All these links are extracted to further iterate through them. Each link is accessed and the HTML of this page is stored in *article*. Further, the function *validateArticle*() checks if the article is listed in the "Wirtschaft" - section. The headline and the actual text of the article are extracted in line 10 and 11. The date, headline and the text are then stored in a database.
The extraction of the specific parts of the HTML document is accomplished by the Beautiful Soup HTML parser.[3] The parser allows to navigate through the Document Object Model (DOM) of the HTML document.

[3] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`, last accessed 24.03.2019

ATX Data	
Date	Closing Price
02.01.2007	4558.96
03.01.2007	4565.91
04.01.2007	4515.17
05.01.2007	4421.55
08.01.2007	4381.50
09.01.2007	4405.16

Table 5.1: Structure of ATX-Data

5.1.2 ATX-Quotation Data

The data describing the historical development of the ATX can be downloaded from the Vienna stock exchange website.[4] Daily measurements about the opening price, the daily maximum, the daily minimum, the closing price and the development in percentage are provided. For the purpose of this thesis the date and the closing price are extracted. The time series of the ATX consist of missing dates, as noticeable in Table 5.1 . Trading opens on Monday and closes on Friday. Further, Austrian holidays are days without trading. The time series, starting from 01.01.2007 until 31.12.2017, consists of 2728 observations where trading took place.

5.2 Sentiment Analysis

This sections explains how the news articles are further processed and analyzed. Furthermore, we describe the preprocessing of the articles and illustrate this process with an example.

5.2.1 Approach

So far the text and the date of issue of the news articles are stored in the database. For the further analysis we group the articles by their publication date. For each day we appended the newspaper articles to retrieve a time series of articles since the data of the ATX is as well measured on a daily basis. Trading on the stock market exchange does not take place on weekends and Austrian holidays. Therefore, only articles published on a date on which trading took place are further considered in order to utilize them to build a model to describe the ATX observations.
Prior to perform the sentiment analysis we preprocess the articles according to the process described by Pröllochs et al. [PFN15], explained in the **Background** chapter. The actual sentiment analysis is accomplished via the lexicon-based approach. The lexicon is

[4] `https://www.wienerborse.at/indizes/aktuelle-indexwerte/historische-daten/?ID_NOTATION=92866&ISIN=AT0000999982`, last accessed 16.08.2018

Sentiment weights					
Date	$Word_1$	$Word_2$	...	$Word_{n-1}$	$Word_n$
d_1	w_1	w_2	...	w_{n-1}	w_n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_t	w_1	w_2	...	w_{n-1}	w_n

Figure 5.1: Post Weighting Structure of the Newspaper Articles

used to capture the sentiments in the preprocessed articles. The lexicon as well as the articles consist of stemmed words. With the GFSL lexicon the articles, published on a trading day, are analyzed. This means for every trading date there exist an observation consisting of the weighted sentiment words of the lexicon. How we perform the weighting of the words is explained in the **Weighting** section. Figure 5.1 shows the weights and the representation of the news articles at a specific date. The words $Word_i$ represent the words in the lexicon. The weights w_i represent the weights of $Word_i$ of the news articles at the specific date d_t. The weights only represent the score computed according to the weighting method. They do not allow for conclusions on the sentiments communicated at this day. They are used to build a supervised machine learning model.

5.2.2 Sentiment Analysis Algorithm

The sentiment analysis of the articles is split into two parts. First, the articles are preprocessed as described in the **Background** chapter. In the second part the weights are computed, according to the weighting schemas, and stored. The following two sections explain these two steps. The functionality of the algorithms to perform these steps is displayed in an abstract form and explained.

Preprocessing

First the dates on which trading took place are extracted by reading the file which contains observations of the ATX-course. The next step is to load the database with the scraped newspaper articles and set up a new database to further store the preprocessed articles. We load the stopwords for the German language. The stopwords and the stemmer are received from the Natural Language Toolkit (NLTK).[5] The NLTK - package uses the Snowball stemming algorithm for German language.[6] For every date, of the previously loaded dates, the newspaper articles are extracted. The articles published on the same date are appended and returned as text. This text is then tokenized with the $WordPunctTokenizer$ from the NLTK package. It splits the text into words and punctuations. The following example illustrates the functionality of the $WordPunctTokenizer$.

[5] `https://www.nltk.org`, last accessed 18.08.2018
[6] `http://snowballstem.org`, last accessed 18.08.2018

Algorithm 2 Preprocessing of the Articles

```
dates ← getDates()
db ← loadDB()
db_processed ← setUpDB()
stopwords ← stopwords.words(”german”)
for each date in dates do
    text ← getArticles(date, db)
    tokens ← WordPunctTokenizer().tokenize(text)
    tokens_processed ← []
    for each word in tokens do
        if word.isalpha() AND word not in stopwords then
            tokens_processed.append(stem(word))
        end if
    end for
    writeProcessedTokens(tokens_processed, date, db_processed)
end for
```

Input:
"An der Wiener Börse stieg der wichtigste österreichische Aktienindex am späten Freitagvormittag um über 1,3 Prozent bis auf 3.003,78 Punkte."

Output:
['An', 'der', 'Wiener', 'Börse', 'stieg', 'der', 'wichtigste', 'österreichische', 'Aktienindex', 'am', 'späten', 'Freitagvormittag', 'um', 'über', '1', ',' , '3', 'Prozent', 'bis', 'auf', '3', '.', '003', ',', '78', 'Punkte', '.']

The algorithm further iterates through every word in the resulting list of words. List elements which consist of non alphabetic characters and punctuations are removed. The remaining words are stemmed and stored in a database. The continuation of the example displays the outcome after the preprocessing step.

Input:
['An', 'der', 'Wiener', 'Börse', 'stieg', 'der', 'wichtigste', 'österreichische', 'Aktienindex', 'am', 'späten', 'Freitagvormittag', 'um', 'über', '1', ',' , '3', 'Prozent', 'bis', 'auf', '3', '.', '003', ',', '78', 'Punkte', '.']

Output:
['wien', 'bors', 'stieg', 'wichtig', 'osterreich', 'aktienindex', 'spat', 'freitagvormittag', 'prozent', 'punkt']

Figure 5.2: Wordcloud Feature Occurrences GFSL

Figure 5.2 displays the feature occurrences of the GFSL in the newspaper articles. The size of the words display the frequency. Bigger words indicate the word occurs more often in the articles. It is important to mention that, although some words occur significantly more often than other words, the occurrences do not reflect the actual importance of the features when building a model. The features which are important will be displayed in the **Results and Discussion** chapter.

Weighting

Rekabsaz et al. [RLB+17] present several weighting methodologies for the sentiments. The following weighting schemes are tested.

TC:	$log(1 + tc_{d_i}(t))$
TCIDF:	$TC * log(1 + \frac{\lvert d_i \rvert}{df(t)})$
TF:	$\frac{TC}{\lvert\lvert d_i \rvert\rvert}$
TFIDF:	$TF * log(1 + \frac{\lvert d_i \rvert}{df(t)})$

where $tc_{d_i}(t)$ describes the occurrences of term t in document d_i. $|d_i|$ is the length of the text, e.g. the number of words. $||d_i||$ describes the L2-Norm of the weights of document d_i. $df(t)$ denotes the document frequency of term t, in this context occurrences of term t in all documents, by documents the news articles published on the same date are referred. In other words on how many days, on which trading took place, does term t occur in a newspaper article.

Algorithm 3 Weighting of the Articles

```
1: dates ← getDates()
2: db ← loadDB()
3: sentiments← getSentiments()
4: sentiments_dict ← dict.fromkeys(sentiments, 0)
5: df ← calculateDF()
6: for each date in dates do
7:     tokens ← getTokens(date, db)
8:     wordcount ← Counter(tokens)
9:     for each word in wordcount do
10:        if word in sentiments_dict then
11:            sentiments_dict[word] = wordcount[word]
12:        end if
13:    end for
14:    writeRowToCSV(weighting(sentiments_dict, df))
15:    sentiments_dict ← dict.fromkeys(sentiments, 0)
16: end for
```

We apply each of the weighting schemes to the features. Like in the preprocessing algorithm the first line loads the dates on which trading took place. In line two the database, where the previously processed articles are stored, is loaded. Further, the sentiment lexicon is loaded. To count the occurrences of the sentiments in the articles we create a dictionary. The dictionary basically consists of two lists, the first list contains the words of the lexicon and the second list the corresponding occurrences. The dictionary is initialized with zeros since some words of the lexicon may not be present at every date. The function *calculateDF*() returns the document frequency of all words in the sentiment lexicon as a dictionary. The document frequency is used for the TCIDF and the TFIDF weighting scheme as explained previous in this section. The algorithm further iterates through all dates and loads the tokens, the list of words resulting from the preprocessing step, for each date. Words in this list are then counted resulting in a dictionary of words and their occurrences. This list contains words which are not in the sentiment lexicon. The words in the lexicon and the tokens are both stemmed from the previous steps enabling the comparison and extraction of only the words which are in the sentiment lexicon.

The occurrences of these words are then stored in the *sentiments_dict*, which was initialized at the beginning with zero values. The dictionary *sentiments_dict* now contains the occurrences of the sentiments of all articles published at the date which is being iterated.

Finally, the sentiment weights, considering the four different weighting schemas, are computed and stored in separate files. The occurrences of the *sentiments_dict* dictionary are then set to zero for the next iteration step.

In this section we established the fundamentals for the machine learning part. The newspaper articles have been cleaned, tokenized, stop words and punctuation have been removed and each word is reduced to its stem, which makes it possible to compare the words in the newspaper articles to the words in the sentiment lexicon. Further, the weighting process has been performed for the lexicons, resulting in four data sets for each sentiment lexicon. Together with the ATX observations these files provide the foundation for the machine learning part.

5.3 Defining the Label

The feature set consists of the sentiment weights, currently four different feature sets per lexicon according to the different weighting schemes calculated previously: TC, TSIDF, TF and TFIDF. The corresponding label vector consists of the closing price of the ATX. The rows of the feature set correspond to the rows in the label vector. They match trading days of the ATX between 01.01.2007 and 31.12.2017. Throughout the process the label vector will be varied in order to find the combination which provides the most accurate model. We test the described label vectors in this section. The y_i represent the label entry and the p_i the closing price observations of the ATX. We investigate regression and classification scenarios. The feature set, the sentiment weights, always stays the same. Only the label set is varied. We perform the experiments with each weighting schema. In the following we define several labels for the regression and the classification experiments.

Labels for Regression:

- For the first attempt to find a model, the label vector is defined as the difference between the closing price of the current date and the date before.

$$y_i = p_i - p_{i-1}$$

- Another possible option for the label vector is the change in volatility. Negative or positive statements in the articles may cause volatility changes. The measurement of volatility is adapted from Rekabsaz et al. [RLB+17].

$$v_{[i,i+t]} = \sqrt{\frac{\sum_{l=i}^{i+t}(p_l - \bar{p})^2}{t}}$$

  For t explaining future volatility changes:

$$y_i = v_{[i,i+t]} - v_{[i-t,i]}$$

  For t explaining past volatility changes:

$$y_i = v_{[i-t,i]} - v_{[i-2*t,i-t]}$$

  The parameter t is varied to investigate different spans of volatility changes.

Labels for Classification:

Directions:

Atkins et al. [ANG18] predict closing price movement and volatility directions. The following labels test whether the sentiment weights allow to classify directions of different values. The resulting label values are classified as positive or negative depending on the sign.

- Closing price directions:

$$y_i = p_i - p_{i-1}$$

- Return directions, as defined in the paper of Rekabsaz et al. [RLB+17] and Atkins et al. [ANG18].

$$y_i = r_i - r_{i-1}, \text{ where } r_i = ln(p_i) - ln(p_{i-1})$$

- Volatility directions, using the volatility measurement mentioned in Rekabsaz et al. [RLB+17] and previously described:

$$y_i = ln(v_{[i,i+t]}) - ln(v_{[i-t,i]})$$

  Again the parameter t is varied.

Abnormal values:

Luss and D'Aspremont [LD15] predict abnormal returns. They classify returns exceeding a threshold into abnormal returns and try to predict these cases with news articles. The following classifications are designed similar to their approach.
In this thesis the threshold in order to classify abnormality is the distance from the mean of the observations. The following measurement is used.

$$mean(v) \pm x * STD(v)$$

The x defines the range for normality, everything outside of this range is defined as abnormal. Different values of x are investigated. Furthermore, classifications of two and of three groups are tested. The two groups are the normal values and the abnormal values. The three group labels distinguish between abnormal low values, normal values and abnormal high values. The different settings are tested for the closing price movements, the returns and the volatility changes as defined previously.

5.4 Experiment Design

This section describes our experiment design. The settings for every experiment is the same. The regression and classification scenarios only differ by the method of evaluation and the tested algorithms. The single steps are performed for each label and each weighting scheme.

1. **Loading:** The first step is to load the sentiment weights and the labels.
2. **Preprocessing:** The features are normalized using the L2 normalization. Further, dimensionality reduction is applied to reduce the time needed for training. We test an amount of 100, 400 and 800 features for each scenario.
3. **Fitting the Model:** For the regression cases we test support vector regression (SVR) and random forest regression (RFR). For the classification cases we test SVC and RFC. In order to find the best hyper-parameter setting, grid search with cross-validation is applied for each algorithm to automatically find the best model according to predefined parameter ranges. Grid search automatically executes every combination of parameter setting and returns the best performing parameter space according to the cross-validation score.
4. **Testing:** The best performing model from step 3 is then applied on the testing set. The resulting predictions are used in the next step.
5. **Evaluation Metrics:**
   - Regression: In the regression case the R2-value and the corresponding p-value are calculated.
   - Classification: To compare classifications we establish a baseline with a dummy classifier. This classifier always predicts the most common group of the training set. The difference between the accuracy of the baseline and the accuracy of the predictions indicates whether the model is capable to capture relevant information from the sentiment weights.

We conduct the experiments with the RFC and SVC algorithms for the classification cases and the RFR and SVR algorithms for the regression cases. We use all four weighting schemas for each experiment. Further, we perform each experiment with the GFSL and the SentiWS lexicon to evaluate the GFSL.

CHAPTER 6

Results and Discussion

In this chapter we present the results of the conducted experiments. The tables display the most interesting results. Some of the tested labels do not provide any useful information, they are not further investigated. In Section 6.1 we present the results of the regression experiments, followed by the results of the classification scenarios in Section 6.2. Section 6.3 is about the qualitative analysis. We discuss the results in Section 6.4.

The following list explains the notation and abbreviations used in this chapter.

- Label: The label used for the model.
- SVR: The Support Vector Regression algorithm.
- SVC: The Support Vector Classification algorithm.
- GFSL: The created German Financial Sentiment Lexicon is used for the sentiment analysis.
- SWS: The SentiWS lexicon is used for the sentiment analysis.
- RFR: The Random Forest Regression algorithm.
- RFC: The Random Forest Classification algorithm.
- R2: The R2 score of the testing set multiplied by 100. Stars indicate the significance of the regression coefficients to the model. The different levels for the p-value are displayed as follows: p-value < 0.05: *, p-value < 0.01: ** and p-value < 0.001: ***
- Baseline: The baseline is defined as the accuracy of the predictions of a dummy classifier which always predicts the most frequent class of the training set. All other settings are compared to this baseline. The $+$ and $-$ signs indicate the difference between the accuracy of the experiment and the baseline.

SVR		RFR	
GFSL	SWS	GFSL	SWS
2.44***	0.99*	0.77*	0.48

Table 6.1: R2 * 100 Closing Price Difference

		SVR		RFR	
Direction	# of days	GFSL	SWS	GFSL	SWS
Future	3	0.03	0.14	**0.91***	0.63
	7	0.32	0.51	0.34	**0.76***
	14	0.28	1.08*	0.70*	**1.31****
	30	0.14	0.20	0.66	**1.18***
	91	0.00	0.19	0.49	**1.71****
	182	2.58***	**6.30*****	1.54**	0.54
	365	9.22***	**14.98*****	3.09***	2.86***
Past	3	0.07	**0.97***	0.28	0.69
	7	0.63	**1.40****	0.89*	0.59
	14	0.00	0.21	0.56	**1.06***
	30	0.15	0.26	0.94*	**1.04***
	91	0.90*	**1.97*****	0.47	1.70**
	182	**2.63*****	1.47**	0.78*	0.37
	365	0.01	0.00	0.33	**0.56**

Table 6.2: R2 * 100 Volatility Changes

- Bold values: In each row of the tables one value is displayed in bold. This setting of algorithm and sentiment lexicon achieved the best results among the other settings.

6.1 Results of the Regression

Table 6.1 displays the performed R2 scores, multiplied by 100, of the two algorithms together with the different lexicons. As observable the sentiment weights contain useful information to predict the development of the closing price. Especially, the case with the SVR algorithm and the financial lexicon performed value of 2.44 with significant influence of the sentiment weights. Both algorithms performed better with the financial lexicon. Support vector regression provide better results in this scenario.
The investigation of the volatility changes, Table 6.2, confirms the support vector algorithm performs better than the random forest algorithm in the regression case. What is further observable, the higher the span the better the results, especially when investigating future volatility changes. This might be due to fewer fluctuations with the higher span. It is also noticeable that the SentiWS lexicon outperforms the financial lexicon in the future significant results. The label, investigating volatility shifts of 365 days into the past, performs surprisingly badly. The label is computed using observations

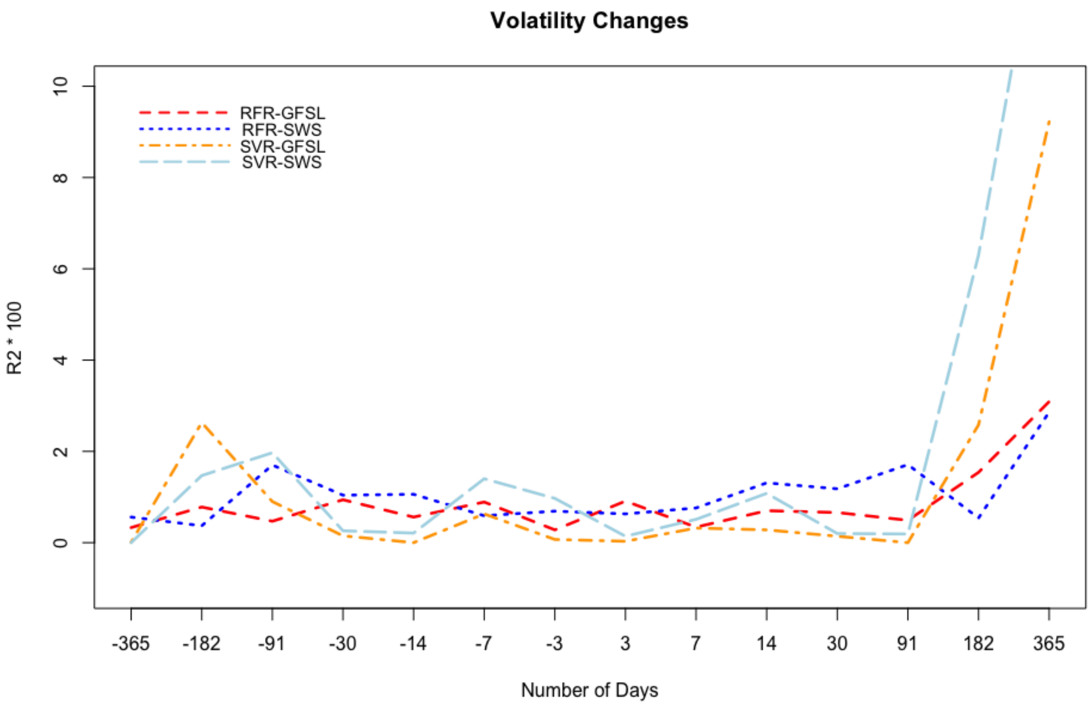

Figure 6.1: R2 * 100 of Volatility Changes based on the results in Table 6.2. The x-axis indicates the span used to compute the volatility.

two years into the past. This means the values are not calculated as intended for the first two years as the data is not available. This explains the fall in performance. Figure 6.1 illustrates the results of Table 6.2. As mentioned the R2 * 100 value significantly increases with spans higher than 91 days into the future.

	SVC		RFC	
Baseline	GFSL	SWS	GFSL	SWS
56.12%	-2.38	-2.74	**+1.28**	-0.37

Table 6.3: Accuracy Closing Price Direction

			SVC		RFC	
Direction	# of days	Baseline	GFSL	SWS	GFSL	SWS
Future	3	49.36%	-1.65	+3.47	**+5.86**	+4.38
	7	51.91%	-0.54	+0.55	+2.93	**+3.11**
	14	53.38%	-1.47	+2.37	**+3.65**	+1.09
	30	46.98%	0.00	0.00	**+3.11**	+1.46
	91	46.25%	0.00	0.00	**+5.48**	+2.74
	182	42.23%	+4.02	+6.39	+5.85	**+6.58**
	365	37.84%	+7.86	**+12.25**	+5.12	+10.97
Past	3	49.36%	+2.58	**+3.47**	+3.29	+3.10
	7	51.55%	0.00	+2.74	**+3.29**	+2.92
	14	53.19%	0.00	0.00	**+0.74**	+0.18
	30	45.70%	0.00	0.00	**+2.92**	+1.46
	91	60.14%	0.00	0.00	**+0.18**	**+0.18**
	182	42.59%	0.00	0.00	0.00	0.00
	365	28.33%	+6.40	+9.93	**+17.55**	+11.52

Table 6.4: Accuracy Volatility Directions

6.2 Results of the Classification

Predicting direction of the closing price resulted in only one model exceeding the baseline, as seen in Table 6.3. The RFC algorithm works best with the GFSL in this test case achieving an increase of 1.28% percentage points over the baseline. All other tested cases are below the baseline.

In contrast to the regression results the classification of volatility changes performs better for shorter time spans, as observable in Table 6.4. The general SentiWS lexicon performs better with the support vector classification algorithm. The random forest algorithm achieved the best results. The GFSL contains words and phrases which are characteristic for positive and negative economic times. The random forest classifier uses these phrases in order to make specific decision rules. This might explain the good results with the financial lexicon, especially in the shorter time spans. We observe a shift in volatility directions. As the baseline always predicts the most frequent group of the training set we can clearly see there is a shift in the test set because the accuracy is lower than 50%. We observe the wider the span the stronger the shift, except the label explaining the volatility directions 91 days into the past. The algorithms are able to detect this shift and perform better results than the baseline. Still these results are

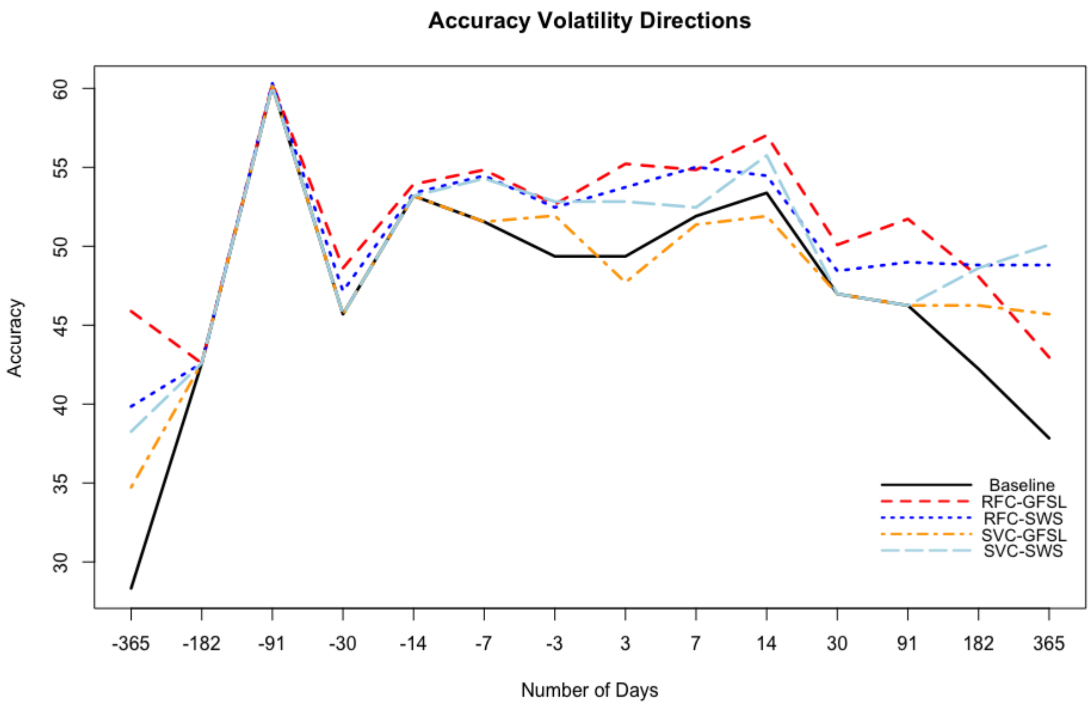

Figure 6.2: Prediction accuracy of Volatility Directions based on the results in Table 6.4. The x-axis indicates the span used to compute the volatility.

often lower than 50% which makes them irrelevant as random classifying would lead to approximately 50% in the long run. The spans up to 14 days into the future and into the past provide satisfying results.

Figure 6.2 illustrates the results presented in Table 6.4. We can see the red line, describing the results of the RFC algorithm with the GFSL, almost has the highest accuracy throughout the spans. As mentioned the predictions work well for short spans into the future and into the past which is observable in the concave trend of all five lines. Table 6.5 describes the results of classifying abnormal values. We investigate different settings for x, regulating the amount of values which are classified as abnormal, ranging from 0.1 up to 2. We achieve the best results with a x value of 0.1 and classification into three groups; abnormal low, normal and abnormal high values. The random forest algorithm, in general, provided better results, despite the cases listed in the table. In our experiments volatility directions with a span of more than 14 days into the past do not provide any useful outcome.

Direction	Algorithm	Label	Baseline	GFSL	SWS
-	SVC	Closing Price	50.09%	**+2.01**	-2.56
	RFC	Volatility 3d.	43.87%	+1.28	**+2.38**
	RFC	Volatility 7d.	46.62%	+0.54	**+2.55**
	SVC	Volatility 14d.	45.52%	+0.91	**+3.10**
Future	RFC	Volatility 30d.	34.73%	**+3.48**	+2.56
	RFC	Volatility 91d.	38.21%	+2.37	**+3.10**
	RFC	Volatility 182d.	36.56%	+3.11	**+3.65**
	RFC	Volatility 365d.	35.10%	**+11.33**	+11.04
	RFC	Volatility 3d.	43.69%	+1.46	**+4.75**
	RFC	Volatility 7d.	46.25%	+1.46	**+3.11**
	RFC	Volatility 14d.	45.52%	**+2.93**	+0.91
Past	RFC	Volatility 30d.	75.13%	0	0
	SVC	Volatility 91d.	70.38%	0	0
	RVC	Volatility 182d.	34.00%	0	0
	RFC	Volatility 165d.	26.14%	0	0

Table 6.5: Accuracy Abnormal Values, with 3 Classes

Figure 6.3: Feature Importance GFSL

Figure 6.4: Feature Importance SWS

6.3 Qualitative Analysis

Figure 6.3 shows the most important features of the classification of the closing price directions using the GFSL with the random forest classifier. Figure 6.4 displays the feature importance with the same setting but using the SWS lexicon for the sentiment analysis. The most important features of the RFC are displayed as the model allows to extract the most important features. The size of the words indicate their importance.

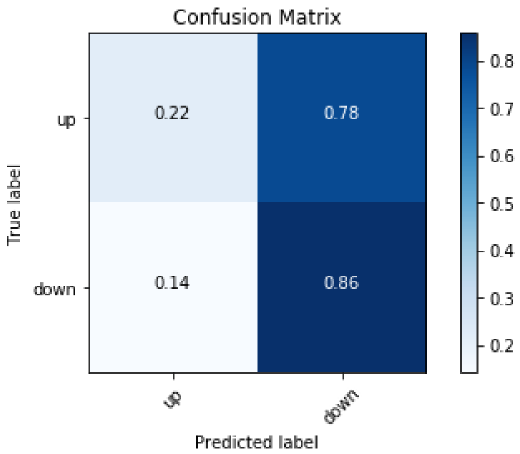

Figure 6.5: Closing Price Difference GFSL

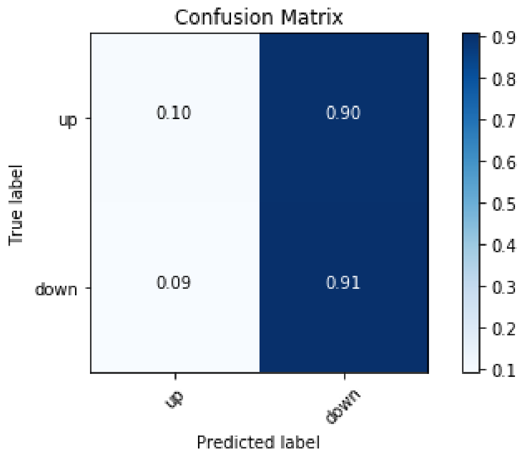

Figure 6.6: Closing Price Difference SWS

The bigger the words the higher their importance in the random forest model. Although the GFSL contains positive and negative associated words, with respect to the economic and financial domain, the words in Figure 6.3, which are immediately recognized, are negatively associated when thinking about a stock market index or an economy. For example "fiel", "verlor" and "tief". The feature importance of the words, derived from the model using the SWS lexicon, Figure 6.4, are more general, which is not surprising as it is the intention of the lexicon to provide a general corpora of German sentiment words.

The confusion matrix of the predicted labels of the RFC algorithm and the true labels, Figure 6.5 for the GFSL and Figure 6.6 for the SWS lexicon, show the predictions are heavily biased towards the negative side, predicting a negative movement of the ATX, for both lexicons. These observations are similar to the ones by Akhtar et al. [AFOS13]. As explained in their work, they find no market effect upon the release of good news while negative news have an impact on the stock market.

The algorithm is able to connect negative sentiments to decreasing movements on the stock market but unable to connect positive sentiments in the newspaper articles to upward movements on the stock market. Further, Schumaker et al. [SZHC12] claim that especially negative subjectivity of news articles have an impact on trading behavior.

We expected the combination of the two lexicons to combine the strengths of the GFSL in some cases and the strengths of the SentiWS in other test cases. This is not the case. Combining the lexicons leads on the one hand to improvements in some experiments but on the other hand to worse results than the other lexicons. There is no clear pattern observable.

6.4 Discussion

This list summarizes the results and main outcomes of this thesis. The following insights are gained:

- In general the sentiments extracted from the newspaper articles contain useful information to build regression and classification models to predict closing price changes, closing price movements, changes in volatility and volatility movements.
- The quality of the predictions varies with the algorithm and lexicon used for the analysis. The sentiments alone do not allow for qualitative predictions. However, they contain useful information which may be used in more advanced models especially when it comes to the prediction of volatility and changes in volatility.
- In general the support vector algorithm performs better when it comes to the regression scenarios while the random forest algorithm provides better results when classifying volatility changes and abnormal values.
- The random forest algorithm performs better with the GFSL in the classification of volatility changes with a small span.
- Regression models, describing volatility changes, are more accurate with higher spans, contrary to classifying volatility changes which work better with a lower span.
- Although the created GFSL does not clearly outperform the SentiWS lexicon, it has advantages in specific test cases. In our experiments the GFSL outperformed the SentiWS in the classification of closing price directions and the prediction of closing price changes.
- The outcomes are highly influenced by negativity in the sentiments. The tested models are able to relate negative sentiments to a decrease of the ATX, positive market behavior is however hardly relatable to positive sentiments.

CHAPTER 7

Conclusion and Future Work

Sentiment analysis of financial news, statements, and text, describing the financial or economic domain, has gained increasing interest of researchers and practitioners.
This thesis focuses on analyzing newspaper articles of an Austrian daily newspaper as a source to predict the development of the Austrian Traded Index (ATX). We recognized a lack of resources to perform sentiment analysis of German financial texts. We therefore contribute to this research area by creating a fin.-specific lexicon, referred to as German Financial Sentiment Lexicon (GFSL), insights in the extraction of daily newspaper articles, and extensive analysis on the possibilities in explaining the behavior of the ATX by the sentiment analysis of news articles.
Predicting the ATX with newspaper articles is a challenging task. We conduct several experiments to study the correlation between the sentiments, extracted from the articles, and the ATX. We train supervised machine learning models to predict closing price changes, volatility changes, closing price movements, volatility movements and abnormal values. We evaluate the models with cross-validation techniques and baselines.
Our experiments show the newspaper articles do contain useful information, which positively correlate with the development of the ATX. Further, we observe the GFSL does not clearly outperform the general SentiWS lexicon, although it has advantages in predicting closing price changes and classifying closing price movements. The outcomes of our experiments are highly influenced by negativity in the sentiments. The tested models are able to relate negativity in the sentiments to negative market behavior while positivity is hardly relatable to positive market behavior confirming the results of other studies.
For future work we consider the following directions:

- The GFSL is one point to improve the results. In our experiments we show the potential of the GFSL. Creating a domain specific sentiment lexicon requires expertise in linguistics. Extending the GFSL with more domain specific words, expressions and language constructs will further increase the quality of future outcomes.

- We recommend to test more machine learning algorithms. An interesting direction is trying Recurrent Neural Networks (RNN), as it considers the development of the sentiments over time. To include a history of sentiments into a model enables to identify phases of positive and negative market behavior.
- This analysis focuses on the articles of one newspaper. We see potential in combining several sources, for example news published by companies listed on the stock exchange as well as using the contents of several online newspapers. Several studies combine sentiment analysis and factual market data to build prediction models. Other economic influences add useful information to such a model and the combination of textual and factual market data enables new possibilities.

List of Figures

List of Tables

List of Algorithms

Bibliography

[AFOS13] Shumi Akhtar, Robert Faff, Barry Oliver, and Avanidhar Subrahmanyam. Stock salience and the asymmetric market effect of consumer sentiment news. *Journal of Banking and Finance*, 37:4488—-4500, 2013.

[ANG18] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120 – 137, 2018. ISSN: 2405-9188.

[BPW19] Christina Bannier, Thomas Pauls, and Andreas Walter. Content analysis of business communication: introducing a german dictionary. *Journal of Business Economics*, 89(1):79–123, Feb 2019.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN: 1573-0565.

[CC17] Samuel W.K. Chan and Mickey W.C. Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64, 2017.

[CDBF17] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A Practical Guide to Sentiment Analysis*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN: 3319553925.

[HCL03] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[HSK18] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura. Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data Science*, 4(3):183 – 201, 2018. ISSN: 2405-9188.

[Jol11] Ian Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN: 978-3-642-04898-2.

[KSR16] Roman Klinger, Surayya Samat Suliya, and Nils Reiter. Automatic Emotion Detection for Quantitative Literary Studies – A case study based on Franz

Kafka's "Das Schloss" and "Amerika". In *Digital Humanities 2016: Conference Abstracts*, pages 826–828, Kraków, Poland, July 2016. Jagiellonian University and Pedagogical University.

[LD15] Ronny Luss and Alexandre D'Aspremont. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012, 2015.

[Liu15] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015. ISBN: 9781139084789.

[MG16] A.C. Müller and S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. ISBN: 9781449369897.

[Mit15] Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 1st edition, 2015. ISBN: 1491910291.

[NH15] Clemens Nopp and Allan Hanbury. Detecting risks in the banking system by sentiment analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 591—-600, 2015.

[PFN15] Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. Enhancing sentiment analysis of financial news by detecting negation scopes. In *2015 48th Hawaii International Conference on System Sciences*, pages 959 – 968. IEEE, 2015. ISBN: 9781479973675.

[Ras15] Sebastian Raschka. *Python Machine Learning*. Packt Publishing, 2015. ISBN: 9781783555130.

[RLB+17] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Andersson. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1712–1721, 2017.

[RQH10] R. Remus, U. Quasthoff, and G. Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, 2010.

[SC09a] Robert P. Schumaker and Hsinchun Chen. A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5):571 – 583, 2009. ISSN: 0306-4573.

[SC09b] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfintext system. *ACM Transactions on Information Systems*, 27, 2009.

[SS04] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. ISSN: 0960-3174.

[SZHC12] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53:458—-464, 2012.

[Tet07] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62:1139–1168, 2007.

[Wal10] Ulli Waltinger. Germanpolarityclues: A lexical resource for german sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010. electronic proceedings.

Attachments

GFSL

abbau
abbrech
abbruch
abfall
abgab
abgemacht
abgesichert
abgesturzt
abgewertet
abhang
ablehn
abmach
abmachet
abmacht
abmachtet
abn
abnahm
abnehm
abrutsch
abschaff
abschliess
abschluss
abschreck
abschwach
absenk
absich
absicher
abstieg
abstoss
absturz
absturzet
absturzs
absturzt
absturztet
absurd
abwartstr
abweich
abwert
abwertet
abwertetet
abwertt
abwerttet
abzeptabl
abzock
affar
aggression
aggressivitat
agil
agilitat
akquisition
aktiv
aktivitat
aktivst
akzeptabel
akzeptabl
akzeptanz
akzepti
akzeptiert
alarm
alarmbereitschaft
alptraum
andrang
androh
anfall
angebot
angekurbelt
angemess
angemessen
angemessn
angenahert
angereichert
angeseh
angesehen
angesehn
angespannt
angriff
angstlich
anheb
ankurbel
ankurbeln
ankurbelt
ankurbeltet
ankurbl
anlag
anleg
annah
annaherst
annahert
annahertet
annehmbar
annehmbarst
annullier
anpass
anpassungsfah
anreich
anreicher
anreicherst
anreichert
anreichertet
anreiz
ansehn
anspann
anspannet
anspannt
anspanntet
ansprech
anspruch
anspruchsvoll
ansteig
anstieg
anstreng
anteil
antreib
antrieb
arbeitslos
arbeitslosst
arm
armst
armut
atemberaub
attraktiv
attraktivitat
attraktivst
aufbess
aufbesser
aufbesserst
aufbessert
aufbessertet
aufgebessert
aufgebracht
aufgehob
aufgemuntert
aufgeregt
aufgestockt
aufgewertet
aufheb
aufhebt
auflos
aufmunt
aufmunternd
aufmunterst
aufmuntert
aufmuntertet
aufreg
aufreget
aufregt
aufregtet
aufruhr
aufschrei
aufschreis
aufschwing
aufschwung
aufsteig
aufstieg
aufstock
aufstocket
aufstockt
aufstocktet
auftrag
auftrieb
aufwand
aufwart
aufwartstr
aufwend
aufwert
aufwertet
aufwertetet
aufwertt
aufwerttet
ausbau
ausbauet
ausbaust
ausbaut
ausbautet
auseinandersetz
ausfall
ausfuhr
ausgab
ausgebaut
ausgeg
ausgeglichen
ausgeglichn
ausgeschaltet
ausgeweitet
ausgewog
ausgewogen
ausgewogn
ausgezeichnet
ausgleich
ausreich
ausschalt
ausschaltet
ausschaltetet
ausschaltt
ausschalttet
ausscheid
ausserordent
aussichtslos
aussteig
austeigt
ausweit
ausweitet
autonomi
auweitet

bahnbrech
bankrott
barri
beachtenswert
beachtlich
beauftrag
beauftraget
beauftragt
beauftragtet
bedarfsgerecht
bedarfsorientiert
bedau
bedenk
bedeut
bedeutsam
bedroh
bedrohet
bedroht
bedrohtet
beeindruck
beeindrucket
beeindruckt
beeindrucktet
beeintracht
beeintrachtiget
beeintrachtigt
beeintrachtigtet
beend
beendet
beendetet
beendt
beendtet
beford
beforder
beforderst
befordert
befordertet
befried
befriediget
befriedigt
befriedigtet
befrist
befristet
befurcht
befurchtet
befurchtetet
befurchtt
befurchttet
begehrt
begeist
begeister
begeisterst
begeistert
begeistertet
begrenz
begrenzet
begrenzt
begrenztet
begunst
begunstiget
begunstigt
begunstigtet
behag
beherrscht
behind
behinderst
behindert
behindertet
beigesteuert
beileg
beisteu
beisteuerst
beisteuert
beisteuertet
beitrag
beitret
beitritt
bekannt
bekraft
bekraftiget
bekraftigt
bekraftigtet
belast
belastbar
belastbarst
beleb
belebet
belebt
belebtet
beliebt
belohn
belohnet
belohnt
belohntet
bemerkenswert
bemuh
bemuht
berausch
berechn
berechnet
bereich
bereicher
bereicherst
bereichert
bereichertenbereit
bereichertet
bereit
bericht
beruh
beruhiget
beruhigt
beruhigtet
beruhmt
beschad
beschadigt
beschaft
beschaftigt
beschaftigungslos
beschaftigungslosst
bescheid
bescheiden
bescheidn
beschleun
beschleuniget
beschleunigt
beschleunigtet
beschrank
beschranket
beschrankt
beschranktet
beschuld
beschuldigt
beschw
beschwerd
beseit
beseitiget
beseitigt
beseitigtet
besitz
besitzs
besond
besonder
besondere
besonderer
besondererst
besorgnis
besorgniserreg
besorgt
bess
besser
besserst
bessert
bessertet
besserungn
best
bestat
bestatiget
bestatigt
bestatigtet
bestmog
bestraf
bestrafet
bestraft
bestraftet
beteil
beteiliget
beteiligt
beteiligtet
betracht
betreff
betrifft
betroff
betrog
betrug
betrugt
beunruh
beunruhiget
beunruhigt
beunruhigtet
bewahrt
bewill
bewilligt
bewirk
bewirket
bewirkt
bewirktet
bewund
bewundernswert
bewunderst
bewundert
bewundertet
bezahl
bezahlt
blendend
blockad
blocki
blockier
bluh
bluhet
bluht
bluhtet
bonitat
bonus
bonuss
boom
breitgefachert
brockeln
brockelnd
brockelt
bussgeld
chaos
chaotisch
comeback
crash
dampf
dampfet
dampft
dampftet
dank
dankbar
dankbarst
danket
dankt
danktet
dauerhaft
deal
defekt
defizit
defizitar
defizitarst
depression
depressiv
depressivst
desast
deutlich
diplomat
diskrediti
diskreditier
diskreditieret
diskreditierst
diskreditiert
diskreditiertet
diss
dissens
disziplin
diszipliniert
divid
dominanz
domini
dominier
dominieret
dominierst
dominiert
dominiertet
drama
dramat
drang
dranget
drangt
drangtet
drastisch
droh
drohet
droht
drohtet
drohung
drossel
drosseln
drosselt
drosseltet
drossl
druck
druckend
druckt
durchdacht
durchhalt
durchschlag
dynam
effektiv
effektivst
effektvoll
effizient
effizienz
ehrgeiz
eif
eiferst
eifert
eifertet
eifrig
eigenkapital
eign
eignet

eignetet
eignt
eigntet
einbrech
einbruch
einbuss
einbusst
eindrucksvoll
einflussreich
eingekauft
eingeschrankt
eingespart
eingesturzt
einhalt
einheit
einkassi
einkauf
einkaufet
einkauft
einkauftet
einkomm
einnahm
einschrank
einschranket
einschrankt
einschranktet
einschuchter
einsink
einspar
einsparet
einsparst
einspart
einspartet
einsteig
einsturz
einsturzet
einsturzs
einsturzt
einsturztet
eintrag
einwandfrei
einwandfreist
energ
engagement
engagi
engagier
engagieret
engagierst
engagiert
engagiertet
enorm
enthusiasmus
enthusiast
entlass
entlast
entlastet
entlastetet
entlastt
entlasttet
entlohn
entschad
entschloss
entschlossen
entschlossn
entschluss
entspann
entspannet
entspannt
entspanntet
enttausch
enttauschet
enttauscht
enttauschtet
entwirr
entwirret
entwirrst
entwirrt
entwirrtet
erarbeit
erarbeitet
erarbeitetet
erarbeitt
erarbeittet
erb
erbet
erbst
erbt
erbtet
erdruck
erfahr
erfahren
erfahrn
erfolg
erfolgreich
erfolgserlebnis
erfreu
erfreuet
erfreulich
erfreust
erfreut
erfreutet
ergebnislos
ergebnisreich
ergieb
erheb
erhoh
erhohet
erhoht
erhohtet
erhol
erholet
erholsam
erholt
erholtet
erklimm
erklimmt
erleicht
erleichter
erleichterst
erleichtert
erleichtertet
erlos
erloset
erlost
erlostet
ermog
ermoglichet
ermoglicht
ermoglichtet
ermunternd
ermut
ermutiget
ermutigt
ermutigtet
erneu
erneuerst
erneuert
erneuertet
ernst
ernuchter
ernuchternd
erreich
erreichet
erreicht
erreichtet
erschutt
erschutternd
erschutterst
erschuttert
erschuttertet
ersparnis
erstaun
erstaunet
erstaunt
erstauntet
erstklass
erstrebenswert
ertrag
ertragreich
erweit
erweiter
erweiterst
erweitert
erweitertet
erwunscht
erzeug
erzeuget
erzeugnis
erzeugniss
erzeugt
erzeugtet
erziel
erzielet
erzielt
erzieltet
etabliert
euphor
euphori
existenzbedroh
exklusiv
exklusivst
exorbitant
expandi
expandier
expandiert
expandirt
expansion
expansionskur
explosiv
explosivst
export
exporti
exportiert
exzellent
fabelhaft
fahig
fahrlass
fall
fallend
fallt
falsch
faszini
faszinier
faszinieret
faszinierst
fasziniert
fasziniertet
fatal
fehl
fehlend
fehleranfall
fehlerfrei
fehlerfreist
fehlerhaft
fehlermeld
fehlkauf
fehltritt
fehlverhalt
feier
feieret
feiern
feierst
feiert
feiertet
fertig
fertiget
fertigt
fertigtet
fest
fiel
finanzi
finanzier
finanzierungskost
finanzkris
finanzspritz
fit
fitness
fitt
flaut
fleiss
fleissig
flexibel
flexibilitat
flexibl
fliess
fliessend
fliesst
flori
florier
florieret
florierst
floriert
floriertet
fluch
folgenschw
folgenschwer
folgenschwerst
ford
forder
forderst
fordert
fordertet
fortdauernd
fortschreit
fortschritt
fragil
fragwurd
freiheit
freu
freud
freudig
freuet
freundlich
freust
freut
freutet
friedlich
frist
fruchtbar
fruchtbarst
frust
frustration
frustri
frustrier
frustrieret
frustrierst
frustriert
frustriertet
fuhr
fuhrend
fuhret
fuhrst

fuhrt
fuhrtet
fuhrung
funktioni
funktionier
funktionieret
funktionierst
funktioniert
funktioniertet
funktionsfah
furcht
furchtbar
furchtbarst
furchterreg
furchtet
furchtetet
furchtlos
furchtt
furchttet
garanti
garantienlgarantiert
garantiert
gebessert
gebluht
gebuhr
gedampft
gedankt
gedrangt
gedroht
gedrosselt
gedruckt
geduld
geeifert
geeignet
geerbt
gefahr
gefahrd
gefahrdet
gefahrdetet
gefahrdt
gefahrdtet
gefall
gefeiert
gefertigt
gefestigt
gefordert
gefreut
gefuhrt
gefurchtet
gegluckt
gehaltszulag
gehemmt
gehofft
gejubelt
gekippt
geklagt
geklettert
gekostet
gekundigt
gekurzt
gelahmt
gelass
gelassen
gelassn
geldgeb
geldstraf
gelernt
gelindert
gelohnt
gelung
gelungen
gelungn
gemeinschaft
gemindert
genehm
genehmiget
genehmigt
genehmigtet
genotigt
genutzt
gerecht
gerechtfertigt
gerettet
gering
gerutscht
geschafft
geschaft
geschaftsauflos
geschick
geschrumpft
geschuldet
gesenkt
gesichert
gesorgt
gespart
gespendet
gestarkt
gesteigert
gestieg
gestockt
gestort
gestottert
gestrahlt
gestreikt
gesturzt
gesunk
getrennt
getrostet
gewach
gewachs
gewackelt
gewagt
gewahrleist
gewalt
gewarnt
gewinn
gewinnbring
gewinnzon
gezielt
gezittert
gezogert
glaubwurd
gleichwert
gluck
gold
golden
goldig
goldn
grandios
gratulation
gravier
greifbar
greifbarst
grenz
grenzenlos
gross
grossart
grossspur
grosst
grosstmog
grosszug
grundleg
grundlich
gunstig
gut
gutst
haft
handel
handeln
handelsaufnahm
handelt
harmon
harmonisi
harmonisier
harmonisieret
harmonisierst
harmonisiert
harmonisiertet
hart
hartnack
haushaltdefizit
haushaltsdefizit
heb
hebt
heftig
heikel
heikl
heit
heiter
heiterst
heitr
hektik
hektisch
hellig
hemm
hemmet
hemmt
hemmtet
hemmung
herabgesetzt
herabsetz
herabsetzet
herabsetzt
herabsetztet
heranwachs
heraufsetz
herausford
herausforder
herausforderst
herausfordert
herausfordertet
herausgefordert
herausrag
herausraged
hervorrag
highlight
hilf
hilflos
hilfreich
hilfsprogramm
hilfszahl
hind
hindernis
hinderniss
hindert
hinreich
hinzufug
hinzufuget
hinzufugt
hinzufugtet
hinzugefugt
hoch
hochattraktiv
hochgestellt
hochgrad
hochklass
hochrang
hochstmog
hochststand
hochtreib
hochwert
hoff
hoffet
hoffnung
hoffnungslos
hoffnungsvoll
hofft
hofftet
hoh
hohepunkt
hundertprozent
ideal
idyll
illegal
imm
immens
imponier
import
importi
importiert
imposant
imspirier
individuell
ineffizient
ineffizienz
inflation
inkompetenz
innovation
innovativ
innovativst
inpirier
insolvent
insolvenz
inspiri
inspirier
instabil
instabilitat
intakt
intelligent
intensiv
intensivst
interess
interessant
interessi
interessier
interessieret
interessierst
interessiert
interessiertet
investor
investi
investier
investieret
investierst
investiert
investiertet
investition
involvi
involviert
jahresgewinn
jahresverlust
jubel
jubeln
jubelt
jubeltet
jubl
kapital
kapitalerhoh
kapitalertragssteu
kapitalverbrech
kapitulation
kassi

kassiert
katastroph
katastrophal
kauf
kauft
kein
kipp
kippet
kippst
kippt
kipptet
klag
klaget
klaglos
klagt
klagtet
klar
klarst
klett
kletterst
klettert
klettertet
knack
knackt
knapp
knappheit
knappst
kollabi
kollabier
kollabieret
kollabierst
kollabiert
kollabiertet
kollaps
kollidi
kollidier
kollidieret
kollidierst
kollidiert
kollidiertet
kollision
komfortabel
komfortabl
kompatibilitat
kompensation
kompensi
kompensier
kompensieret
kompensierst
kompensiert
kompensiertet
kompetent
kompetenz
komplett
kompromiss
konflikt
konfronti
konfrontiert
konjunkturabschw
konjunkturaufschw
konjunkturaufschwung
konjunkturruckgang
konkret
konkur
konkurrenz
konkurrenzfah
konkurrenzkampf
konkurri
konkurs
kons
konsens
konsequent
konsistent
konsistenz
konsolidi
konsolidier
konsolidieret
konsolidierst
konsolidiert
konsolidiertet
konstant
konstanz
konstruktiv
konstruktivst
konsulti
konsultier
konsultieret
konsultierst
konsultiert
konsultiertet
konsum
konsumi
konsumiert
konterproduktiv
kontinui
konzertiert
kooperation
kooperativ
kooperativst
kooperi
koordini
koordiniert
korrekt
korrektur
korrigi
korrigier
korrigiert
korrupt
korruption
kost
kostbar
kostbarst
kostengunst
kostenintensiv
kostenintensivst
kostenlos
kostenlosst
kostet
kraft
kraftig
kraftvoll
kris
kritik
kritisch
kundig
kundiget
kundigt
kundigtet
kurseinbruch
kursgewinn
kursverli
kursverlier
kursverlust
kursziel
kurz
kurzet
kurzt
kurztet
kurzung
labil
lahm
lahmet
lahmt
lahmtet
lahmung
langsam
last
laun
lebensfah
lebhaft
leer
legal
legalitat
leicht
leichtglaub
leichtig
leichtsinn
leid
leidend
leidet
leistung
leistungsfah
leistungsstark
lern
lernet
lernt
lerntet
limit
limiti
limitier
limitiert
lind
linderst
lindert
lindertet
liquid
liquidation
liquiditat
lob
lock
locker
lockerst
lockr
lohn
lohnend
lohnet
lohnsteu
lohnt
lohntet
loyalitat
lukrativ
lukrativst
luxorios
luxurios
macht
machtig
mag
mager
magerst
magr
makellos
mangel
mangelbehaftet
mangelhaft
mangelnd
manipulation
manipuli
marod
massiv
massivst
maximal
maximi
maximier
maximieret
maximierst
maximiert
maximiertet
maximum
meist
meisterhaft
mild
milder
mildernd
mildert
mind
minder
minderst
mindert
mindertet
minus
mis
miserabel
miserabl
misstrau
mittelmass
mobil
mobilisi
mobilisier
mobilisieret
mobilisierst
mobilisiert
mobilisiertet
mobilitat
mod
modern
modernisi
modernisier
moglich
monopol
moral
motivation
motivi
motiviert
muhelos
muhsam
mut
nachfrag
nachhalt
nachlass
nachteil
naiv
naivitat
negativ
negativbescheid
negativitat
negativst
nennenswert
nervos
nervositat
neuordn
nicht
niedergang
niederlag
niedrig
not
notfall
notig
notiget
notigt
notigtet
notstand
nutz
nutzet
nutzlich
nutzt
nutztet
nzureich
offensiv
ohn
optimalitat
optimismus
optimist

ordent
ordnungsgemass
panik
panisch
pann
paradi
paradies
paradiess
partn
partnerschaft
pech
perfektion
perfektionismus
perfektionist
pessimismus
pessimist
planlos
planlosst
planmass
planvoll
pleit
plus
poplaritat
popular
popularitat
positiv
positivitat
potent
potenz
prachtig
prachtig|
prachtvoll
praferenz
praktikabel
praktikabl
praktisch
prami
prazision
preisgunst
preissturz
preissturzs
prekar
prekarst
privilegiert
probl
problem
problemat
problemlos
problemlosst
produktiv
produktivitat
produktivst
professionell
profit
profitabel
profitabl
profiti
profitiert
progression
projekt
prot
protesti
protestier
protestiert
ptimal
qualifikation
qualifizi
qualifizier
qualifizieret
qualifizierst
qualifiziert
qualifiziertet
qualitat
qualitativ
qualitativst
qualitatsverbesser
quartalsgewinn
quartalsverlust
rach
rachend
racht
rational
realisi
realisiert
realist
recht
rechtfert
rechtfertiget
rechtfertigt
rechtfertigtet
rechtlich
rechtmass
rechtsgult
rechtswidr
reduktion
reduzi
reduzier
reduzieret
reduzierst
reduziert
reduziertet
regulation
reguli
regulier
reguliert
reibungslos
reich
reichhalt
reichlich
reichtum
reinfall
reizend
reizvoll
rekord
rekordergebnis
rekordhoch
rekordtief
rekordverlust
rekordwert
relevant
relevanz
rendit
renovier
rentabel
rentabilitat
rentabl
reparatur
respektabel
respektabl
respekti
respektier
respektvoll
respsrektier
rett
rettet
rettetet
rettung
revanchi
revanchier
revanchieret
revanchierst
revanchiert
revanchiertet
rezession
richtig
riesengross
riesig
risiko
riskant
riski
riskier
riskieret
riskierst
riskiert
riskiertet
rivalitat
robust
rosig
ruckgang
rucklauf
ruckschlag
ruckstell
rucktritt
ruh
ruhig
ruin
ruini
ruinier
ruinos
rutsch
rutschet
rutscht
rutschtet
sani
sanier
saniert
sank
schad
schadig
schadlich
schaff
schaffet
schafft
schafftet
scheit
scheiternd
scheitert
schenkung
schlecht
schliessung
schlimm
schmelz
schmelzend
schmelzt
schmolz
schnappch
schnell
schnellig
schock
schockier
schockn
schonungslos
schrumpf
schrumpfet
schrumpft
schrumpftet
schub
schuld
schuldet
schuldetet
schuldhaft
schuldig
schuldlos
schuldn
schuldt
schuldtet
schutz
schutzmassnahm
schutzs
schwach
schwachung
schwer
schwerfall
schwerst
schwerwieg
schwierig
schwung
sehr
senk
senket
senkt
senktet
senkung
sensation
sensationell
sensibel
sensibl
serios
seriositat
sich
sicher
sicherst
sichert
sichertet
sichr
simpel
simpl
sink
sinnlos
sinnvoll
skandal
skandalos
skeptisch
solid
sonnig
sorg
sorgenfrei
sorgenfreist
sorget
sorgfalt
sorglos
sorgsam
sorgt
sorgtet
souveran
spar
sparet
sparkur
sparkurs
sparprogramm
sparsam
sparst
spart
spartet
spektakular
spektakularst
spekulation
spekuli
spekuliert
spend
spendet
spendetet
spendt
spendtet
spitzenverdi
sprung
spurbar
spurbarst
stabil
stabilisi
stabilisier
stabilisieret

stabilisierst
stabilisiert
stabilisiertet
stabilitat
stagnation
stagni
stagniert
standhaft
stark
starket
starkt
starktet
steierhinzerzieh
steig
steigend
steiger
steigerst
steigert
steigertet
stellenabbau
stellenabbaus
stetig
steu
steuererhoh
steuerhinterzieh
stieg
stiftung
stillleg
stillstand
stimuli
stimulier
stimulieret
stimulierst
stimuliert
stimuliertet
stock
stocket
stockt
stocktet
stor
storet
storst
stort
stortet
storung
storungsfrei
storungsfreist
stott
stotterst
stottert
stottertet
straf
strafbar
strafbarst
strafverfahr
strafzins
strahl
strahlend
strahlet
strahlt
strahltet
strapaz
streich
streicht
streik
streiket
streikt
streiktet
strukturi
strukturier
sturm
sturmisch
sturz
sturzet
sturzs
sturzt
sturztet
stutz
stutzend
stutzt
stuzt
subvention
subventioni
subventionier
subventionieret
subventionierst
subventioniert
subventioniertet
tatkraft
tauglich
tendenz
tendi
tendiert
teu
teuerst
teufelskreis
teufelskreiss
teur
tief
tilg
tilgt
tilgung
trag
trage
tragheit
traum
trauma
traumat
traumatisi
trenn
trennet
trennt
trenntet
triumph
triumphi
triumphier
triumphieret
triumphierst
triumphiert
triumphiertet
trost
trostet
trostetet
trostt
trosttet
trub
trubend
trubt
trugschluss
turbolenz
turbulent
uber
uberdurchschnitt
ubergeholt
ubergluck
uberhol
uberholet
uberholt
uberholtet
uberlastet
uberrasch
uberrascht
uberschreit
uberschreitet
ubersteh
ubersteht
ubersteig
ubersteigt
ubertreff
ubertrifft
ubertroff
umbruch
umfangreich
umgang
umstritt
umstritten
umweltschad
unabhang
unangefocht
unangefochten
unangenehm
unantastbar
unantastbarst
unattraktiv
unattraktivitat
unaufholbar
unbefried
unbefristet
unbegrenzt
unbeliebt
unbeschwert
unbesiegbar
unbesiegbarst
unbezahlbar
unbezahlbarst
unbrauchbar
undurchsicht
unein
uneingeschrankt
unerbitt
unerklar
unermess
unermud
unerreicht
unerschrock
unerschrocken
unerschrockn
unersetz
unertrag
unerwartet
unerwunscht
unfah
unfair
unfairst
unfall
unfehlbar
unfehlbarst
ungeahnt
ungebroch
ungebrochen
ungedeckt
ungerecht
ungeschickt
ungesetz
ungewollt
unglaub
ungleich
ungluck
ungunst
universell
unklar
unklarst
unklug
unkompliziert
unlaut
unlauter
unlauterst
unlautr
unmog
unprofitabel
unprofitabl
unqualifiziert
unrealist
unregelmass
unrentabel
unrentabilitat
unrentabl
unruh
unschlagbar
unschlagbarst
unschon
unschuld
unserios
unsich
unsicher
unsicherst
unsichr
unsolidar
unstet
unt
unterbrech
unterbricht
unterbroch
unterdruck
unterdrucket
unterdruckt
unterdrucktet
unterentwickelt
untergang
untergedruckt
untergeh
unternehmer
unterstutz
untreu
unubertreff
unubertroff
unubertroffen
ununterbroch
ununterbrochen
unverantwort
unvergleichbar
unvergleichbarst
unverhaltnismass
unverhofft
unverkrampft
unvermeid
unverstand
unvorhergeseh
unvorhergesehen
unwicht
unwidersteh
unwirksam
unwirtschaft
unwiss
unzufried
unzufrieden
unzufriedn
unzulass
unzumutbar
unzumutbarst
unzurechnungsfah
unzureich
unzuverlass
vehement
verangst
verangstigt
verantwort
verantwortungsbewusst
verantwortungsvoll
verband
verbess
verbesser

verbiet
verbind
verblind
verbluff
verbot
verboten
verbotn
verbrauch
verbraucht
verbundet
verdacht
verderb
verdi
verdien
verdienet
verdient
verdientet
verdirbt
verdorb
veredel
veredeln
veredelt
veredeltet
veredl
vereinbar
vereinfach
verfall
verfallen
verfalln
verfalsch
verfalschet
verfalscht
verfalschtet
verfehl
verfehlet
verfehlt
verfehltet
vergeb
vergemacht
vergeud
vergeudet
vergeudetet
vergeudt
vergeudtet
vergezogert
vergross
vergrosser
vergrosserst
vergrossert
vergrossertet
vergut
vergutet
verhandel
verhandeln
verhandelt
verhandeltet
verhandl
verhangisvoll
verhangnis
verhangniss
verhangnisvoll
verheer
verkauf
verkauft
verklein
verkleiner
verkleinerst
verkleinert
verkleinertet
verlangsam
verlangsamet
verlangsamt
verlangsamtet
verlass
verli
verlier
verlor
verlust
verlustzon
vermach
vermachet
vermacht
vermachtet
vermind
verminder
verminderst
vermindert
vermindertet
vernunft
verrechn
verrechnet
verring
versag
versagt
verschieb
verschiebt
verschlecht
verschlechter
verschmutz
verschmutzet
verschmutzt
verschmutztet
verschwend
verschwender
verschwendet
verschwendetet
verschwendt
verschwendtet
versich
versicher
versicherst
versichert
versichertet
verspiel
verspielet
verspielt
verspieltet
verstandnis
verstandniss
verstark
verstarket
verstarkt
verstarktet
verstoss
verteid
verteidiget
verteidigt
verteidigtet
verteuer
vertrag
vertrau
vertrauenerweck
vertrauensvoll
vertrauenswurd
verurteil
verurteilet
verurteilt
verurteiltet
verwerf
verwirr
verzicht
verzog
verzoger
verzogerst
verzogert
verzogertet
verzweifel
verzweifeln
verzweifelt
verzweifeltet
verzweifl
viel
vielfalt
vielseit
vielversprech
vielzahl
visionar
volum
voranbring
vorankomm
vorantreib
vorgesorgt
vorsicht
vorsorg
vorsorget
vorsorgt
vorsorgtet
vorteil
vorteilhaft
vorurteil
vorwand
vorwart
vorwurf
vorzeigbar
vorzeigbarst
vorzug
wach
wachs
wachsam
wachsend
wachset
wachst
wachstet
wachstum
wachstumschanc
wackel
wackeln
wackelt
wackeltet
wackl
wag
waget
wagst
wagt
wagtet
warn
warnend
warnt
wegweis
weitgeh
weitlauf
weitraum
weitreich
weitsicht
wenig
wert
wertgeschatzt
wertlos
wertschatz
wertschatzet
wertschatzt
wertschatztet
wertsteiger
wertverfall
wertverlust
wertvoll
wettbewerbswidr
wichtig
widerruf
widerspruch
widerstand
widerstandsfah
widrig
wiederhergestellt
wiederherstell
wiederherstellet
wiederherstellt
wiederherstelltet
wiederstand
willkur
wirksam
wirkungslos
wirkungsvoll
wirtschaft
wirtschaftskris
wohlergeh
wohlhab
wohlstand
wund
zahl
zahlt
zahlung
zahlungsunfah
zahlungsverzoger
zerschlag
zerschlagen
zerstor
zerstorer
zerstoret
zerstorst
zerstort
zerstortet
zielgerichtet
ziellos
zielstreb
zitt
zitterst
zittert
zittertet
zog
zogerst
zogert
zogertet
zoll
zolln
zufried
zufrieden
zufriedenstell
zufriedn
zugelegt
zukauf
zukauft
zukunftsweis
zulag
zulass
zuleg
zuleget
zulegt
zulegtet
zunahm
zunehm
zuruckgeh
zuruckging
zuruckhalt
zuruckzieh
zusammenarbeit
zusammenbrech
zusammenbruch
zusammenhalt
zusatz
zuschlag
zuschuss

zusich
zusicher
zusichernd
zustimm
zuverlass
zuversicht
zuvorkomm
zwangslag
zwangsverkauf
zwecklos
zweifel
zweifelhaft
übereinstimm
überfluss
überhoh
überlast
überleg
überschaubar
überschuss