FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Risk Sentiment Analysis in Banking Supervision

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Clemens Nopp
Matrikelnummer 0826054

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Privatdoz. Dr. Allan Hanbury
Mitwirkung: Univ.Lektor Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Bernhard Kronfellner

Wien, 26.02.2015

(Unterschrift Verfasser) (Unterschrift Betreuung)

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Risk Sentiment Analysis in Banking Supervision

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Clemens Nopp
Registration Number 0826054

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Privatdoz. Dr. Allan Hanbury
Assistance: Univ.Lektor Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Bernhard Kronfellner

Vienna, 26.02.2015

(Signature of Author) (Signature of Advisor)

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Clemens Nopp
Pappenheimgasse 1/16, 1200 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.

(Ort, Datum) (Unterschrift Verfasser)

Danksagung

Zu Beginn möchte ich mich bei Privatdoz. Dr. Allan Hanbury für die hervorragende Betreuung und Unterstützung bei dieser Diplomarbeit bedanken. Insbesondere danke ich ihm für seine Hilfe bei der Themenwahl sowie für die detaillierten und hilfreichen Kommentare zu meinen Entwürfen, welche die Qualität der Arbeit sicherstellten.

Ich möchte mich außerdem bei Univ.Lektor Dr. Bernhard Kronfellner (The Boston Consulting Group) bedanken, der mit mir die initiale Idee für diese Arbeit entwickelt und sich immer gerne Zeit für die Beantwortung meiner finanzspezifischen Fragen genommen hat.

Mein besonderer Dank gilt Johanna Hiebl, BSc (WU), für ihr Verständnis, den Rückhalt während meiner gesamten Studienzeit und das Korrekturlesen meiner Diplomarbeit.

Abschließend möchte ich mich herzlich bei meinen Eltern für ihre Unterstützung und ihren Beistand in allen bisherigen Lebensphasen bedanken.

Wien, im Februar 2015
Clemens Nopp

Abstract

In November 2014, the European Central Bank (ECB) started to directly supervise the largest banks in the Eurozone via the Single Supervisory Mechanism (SSM). Supervisory risk assessments are primarily based on quantitative data and surveys, but textual disclosures as another source of information have remained largely untapped so far. This work utilizes these data by exploring a novel application of sentiment analysis. It evaluates whether this popular approach in the field of text mining is capable of measuring a bank's attitude and opinions towards risk.

For realizing this study, a text corpus consisting of more than 500 CEO letters and outlook sections extracted from annual reports is built up. The documents were published by banks in the Eurozone and cover the period from 2002 to 2014. Based on these documents, two distinct experiments are conducted. The first one derives sentiment scores for measuring the degrees of uncertainty, negativity, and positivity in the documents. The scores are determined based on a finance-specific lexicon and term weighting techniques. Another experiment employs machine learning algorithms for risk classification. The results are evaluated both in a qualitative way and by comparison with Tier 1 capital ratios, which are quantitative risk indicators in the financial industry.

The evaluations find promising opportunities, but also limitations for risk sentiment analysis in banking supervision. At the level of individual banks, it can only inaccurately predict whether the quantitative risk indicator will rise or fall within the following year. In contrast, the analysis of aggregated figures revealed strong and significant correlations between uncertainty or negativity in textual disclosures and the Tier 1 capital ratio's future evolution. Risk sentiment analysis should therefore rather be used for macroprudential analyses than for risk assessments of individual banks. Furthermore, the aggregated sentiment scores clearly reflect major economic events between 2002 and 2014, for example the financial crisis. This facilitates the creation of risk sentiment indicators.

Kurzfassung

Im November 2014 übernahm die Europäische Zentralbank (EZB) im Rahmen des *Single Supervisory Mechanism* (SSM) die Aufsicht über die größten Banken der Eurozone. Risikobewertungen der Bankenaufsicht basieren üblicherweise auf quantitativen Daten sowie Umfragen. Textdaten als weitere Informationsquelle wurden bislang nur unzureichend genutzt. In der vorliegenden Arbeit werden diese Daten im Rahmen eines neuen Anwendungsgebietes der Sentimentanalyse genutzt. Es wird evaluiert, ob diese populäre Methode aus dem Forschungsgebiet *Text Mining* dafür geeignet ist, die Einstellungen und Meinungen von Banken gegenüber Risiken zu messen.

Für die Realisierung der Studie wird ein Textkorpus mit über 500 Vorstandsbriefen und Prognosen aus den Jahresberichten zusammengestellt. Die Dokumente wurden von Banken der Eurozone zwischen den Jahren 2002 und 2014 veröffentlicht. Basierend auf diesen Dokumenten werden zwei voneinander unabhängige Experimente durchgeführt. Ersteres ermittelt Sentimentpunkte, welche die Grade der Unsicherheit, Negativität und Positivität in den Dokumenten messen. Die Ermittlung der Punktzahlen basiert auf einem finanzspezifischen Lexikon und Techniken zur Gewichtung von Worten. Ein weiteres Experiment verwendet Algorithmen des maschinellen Lernens für Risikoklassifizierungen. Die Ergebnisse werden sowohl qualitativ als auch per Vergleich mit Kernkapitalquoten evaluiert. Letztere fungieren als quantitatives Risikomaß im Bankenbereich.

Die Evaluierungen ergeben vielversprechende Möglichkeiten, aber auch Limitationen für den Einsatz von Risiko-Sentimentanalyse in der Bankenaufsicht. Auf der Ebene von individuellen Banken kann nur ungenau vorhergesagt werden, ob der quantitative Risikoindikator im folgenden Jahr steigen oder fallen wird. Im Gegensatz dazu zeigen die Auswertungen der aggregierten Zahlen starke und signifikante Korrelationen zwischen Unsicherheit oder auch Negativität in Textdaten und der Entwicklung der Kernkapitalquote. Risiko-Sentimentanalyse sollte daher eher für makroprudentielle Analysen als für Risikobewertungen einzelner Banken genutzt werden. Darüber hinaus spiegeln die aggregierten Sentimentpunkte wichtige wirtschaftliche Ereignisse im Zeitraum zwischen 2002 und 2014 deutlich wider. Dies ermöglicht die Erstellung von Risiko-Sentimentindikatoren.

Contents

Acronyms

BCBS	Basel Committee on Banking Supervision.
BoW	Bag of Words.
BRRD	Bank Recovery and Resolution Directive.
CDS	Credit Default Swap.
CEO	Chief Executive Officer.
CLIR	Cross-lingual Information Retrieval.
COREP	Common Reporting.
CRD	Capital Requirements Directive.
CRR	Capital Requirements Regulation.
DF	Document Frequency.
DGSD	Deposit Guarantee Scheme Directive.
DPM	Data Point Model.
EBA	European Banking Authority.
EBU	European Banking Union.
ECAI	External Credit Assessment Institution.
ECB	European Central Bank.
ENET	Elastic-Net Logistic Regression.
ER	Entity Relationship.
ESFS	European System of Financial Supervisors.
ESI	Economic Sentiment Indicator.
ESM	European Stability Mechanism.
ESRB	European Systemic Risk Board.

EU	European Union.
FINREP	Financial Reporting.
FN	False Negatives.
FP	False Positives.
FSB	Financial Stability Board.
GATE	General Architecture for Text Engineering.
GDP	Gross Domestic Product.
GI	General Inquirer.
H4N	Harvard IV Negative Word List.
IDF	Inverse Document Frequency.
IE	Information Extraction.
IFRS	International Financial Reporting Standards.
IG	Information Gain.
IR	Investor Relations.
ITS	Implementing Technical Standards.
KRI	Key Risk Indicator.
MD&A	Management's Discussion and Analysis of Financial Conditions and Results of Operations.
MT	Machine Translation.
NB	Naïve Bayes.
NCA	National Competent Authority.
NLP	Natural Language Processing.
OLS	Ordinary Least Squares.
OMSVM	Ordinal Multi-Class Support Vector Machines.
OPP	Ordinal Pairwise Partitioning.
POS	Part-of-Speech.
RAS	Risk Assessment System.
RIAD	Register of Institutions and Affiliates Database.
RWA	Risk-Weighted Asset.
SD	Standard Deviation.
SEC	Securities and Exchange Commission.
SRB	Single Resolution Board.

SRF	Single Resolution Fund.
SRM	Single Resolution Mechanism.
SSM	Single Supervisory Mechanism.
SUBA	Supervisory Banking.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
T1	Tier 1 Capital Ratio.
TF	Term Frequency.
TF-IDF	Term Frequency - Inverse Document Frequency.
TN	True Negatives.
TP	True Positives.
VaR	Value-at-Risk.
VSM	Vector Space Model.
XBRL	eXtensible Business Reporting Language.

CHAPTER 1

Introduction

"The euro area needs healthy banks, which enjoy the confidence of the people and the markets. Only these institutions will be able to finance the real economy and support job creation. The SSM [Single Supervisory Mechanism] will provide a far-reaching and thorough supervisory framework for better early detection of the risks developing within a financial institution"

— Danièle Nouy, *European Central Bank (2014a)*

1.1 Motivation

From 2007 on, a global crisis struck the financial markets and led to a severe slow-down of the real economy. It was triggered by the collapsing US subprime mortgage sector, where loans had been issued to borrowers with poor credit ratings. Due to the tight interconnectedness of the financial system, problems quickly propagated in the global banking system. Governments had to bail out important institutions like *Northern Rock*, but such solutions could not be provided for every troubled bank. In September 2008, the large investment bank *Lehman Brothers* had to file bankruptcy. In the aftermath of this event, further banks had to be rescued in order to stabilize the financial system. This deep financial crisis highlighted the necessity of better financial regulation as well as more effective financial supervision in the future (cf. Hodson and Quaglia, 2009).

In the meantime, the *European Central Bank* (ECB) started to implement a new system of financial supervision, the *Single Supervisory Mechanism* (SSM). Since November 2014, the ECB has been supervising 123 European large banks directly and thousands of other banks indirectly via *national competent authorities*. The SSM will be a part of the banking union in Europe and "confers on the ECB specific tasks concerning policies relating to the prudential supervision of credit institutions, with a view to contributing to the safety and soundness of credit institutions and the stability of the financial system within the Union and each Member State [...]" (Council of the EU, 2013, p. 72).

One the one hand, supervisors collect huge amounts of quantitative data for risk assessments, quality reviews, and peer group analyses. The suggested basis for these supervisory reports is an extensive XBRL taxonomy[1] (European Banking Authority, 2013).

On the other hand, the ECB could supplement these analyses with so-called big data methodologies in order to gain more insights. The ECB is definitely interested in such topics since they held a big data workshop in 2014[2]. Other central banks are dealing with the topic as well. For example, the *Bank of Canada* identifies several opportunities for the use of big data in current analysis (Armah, 2013). They highlight the timeliness of data when employing such approaches, but admit that there are issues with privacy and methodological constraints. The professional services firm *Deloitte* published a document on big data issues in connection with the SSM (Deloitte LLP, 2014). They suggest that supervisors should form a consistent view by utilizing a range of information sources. The latter include not just organized, structured data, but also unstructured information like reports. Deloitte notes that techniques like text analytics could be used for extracting and structuring information from such sources. Furthermore, this technology could be used within portals where banks submit data to supervisors. A more specific example for the tasks of supervisors are risk assessments. In one of its SSM Quarterly Reports, the ECB outlines the objectives of such procedures: "It would embody a quantitative and qualitative analysis based on backward and forward-looking information aimed at assessing a bank's intrinsic risk profile, its position in relation to peers and its vulnerability to a number of exogenous factors" (European Central Bank, 2014c, p. 14f).

These examples document the relevance of further work in this context. In particular, *sentiment analysis* is a promising approach in the field of text analytics. While there is no universal definition of this term, this work uses it as "computational treatment of opinion, sentiment, and subjectivity in text" (Pang and Lee, 2008, p. 6). Academic work in the context of sentiment analysis often tries to classify sentences or documents according to their polarity, for example whether a product review expresses a positive, neutral or negative sentiment (Liu, 2010). There exists much less research focusing on sentiment analysis in the context of business risks. This field focuses on exploring attitudes and opinions about risk and uncertainty in textual company disclosures. Existing literature usually tries to identify relations between risk sentiment and future earnings[3]. Interestingly, there seems to be no work dealing with risk sentiment in the banking sector. This makes it an interesting field to work on in the course of this thesis.

1.2 Problem Definition

Organizations like the *European Banking Authority* (EBA) and supervisors are interested in the financial stability of banks. In an extensive process, they analyze data from the European banking sector in order to detect vulnerabilities and risks. The EBA uses a set of 53 *Key Risk Indicators* (KRIs) for its risk assessments. The mainly backward-looking KRIs are complemented with

[1] XBRL is XML-based and stands for eXtensible Business Reporting Language.

[2] See http://www.ecb.europa.eu/events/conferences/html/20140407_call_international_institute_of_forecasters.en.html, accessed March 21th, 2014.

[3] See for example Groth and Muntermann (2011) or Li (2006).

surveys in order to include forward-looking information as well (European Banking Authority, 2014b).

However, another source of information seems to be largely untapped, namely textual data published by the banks. Publications like periodic reports, press releases, and news published for investors also contain forward-looking information. Analyzing this readily available data would be more cost-efficient in comparison to traditional approaches like surveys. It could provide answers to questions like: what does official communication by banks reveal about their expectations and attitudes towards risk?

In order to answer such questions, a number of challenges have to be tackled. First, proper documents have to be identified and collected. They have to be publicly available, forward-looking and to contain subjective information about risks. Second, these data have to be processed and analyzed. The challenge here is to identify the best suited approach for extracting this information. Finally, the outcome of the analysis has to be interpreted in a meaningful way.

This work contributes to the research field of sentiment analysis in finance by exploring how banking supervisors could utilize text mining for evaluating the risk sentiment of banks respectively their representatives' opinions about the current and future situation.

1.3 Aim of the Work

This thesis answers the following question: to which extent is it possible to measure a bank's attitude and opinions about risk by utilizing sentiment analysis? In this context, the term denotes the automated analysis of documents by banks with the objective to reveal information about their risk sentiment and estimations regarding risk factors in the near future. In order to answer this question, possibilities and limitations of sentiment analysis in banking supervision, especially in the EU's new Single Supervisory Mechanism (SSM), will be explored. This involves answering the questions outlined below.

Which Data Sources are Suitable for Risk Sentiment Analysis? A careful selection of data sources is crucial since irrelevant documents would lead to biased conclusions. On the one hand, documents could be misleading, e.g. if it is unclear to which bank a statement refers. On the other hand, text data could simply add noise if it does not provide any information about risks.

This thesis discusses and evaluates several types of documents which could be used for the analysis. Besides content-related aspects, the data also have to be publicly accessible. Examples for potentially interesting documents are letters to the shareholders by CEOs, press releases, risk reports, or social media postings.

Furthermore, a suitable quantitative risk indicator has to be identified. It needs to represent the bank's financial health and its general risk exposure. Further requirements for such an indicator are comparability and public accessibility.

Which tools and techniques should be used for extracting risk sentiment? There are different approaches for conducting sentiment analysis in textual statements and documents. But which ones are appropriate for risk sentiment analyses in the banking sector? Answering this question involves the analysis of related work, investigations about domain-specific dictionaries

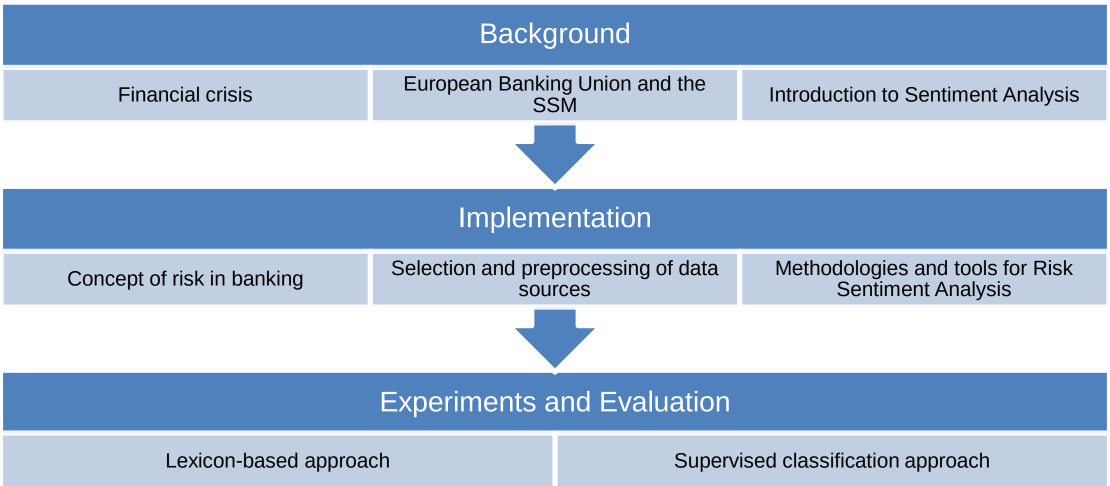

Figure 1.1: Summary of the methodological approach.

and whether machine learning techniques can be used for predicting the evolution of quantitative risk indicators.

How can the results be interpreted and utilized? The outcome of the sentiment analysis reflects a bank's attitudes and opinions about risk. While there are many possibilities for interpreting the figures, the evaluation within this work analyzes the evolution of risk sentiment within the last decade in order to cover the financial crisis of 2007-08 adequately. It is of particular interest if there are correlations with quantitative risk indicators. If risk sentiment in forward-looking statements is a leading indicator for the actual risk figures, it can be used within predictive models.

1.4 Methodological Approach

In order to answer the questions described above, this work is approached according to methodology depicted in Figure 1.1.

First, the necessary background knowledge is built up. This involves a discussion of the financial crisis from the banks' point of view and the EU's stabilizing measures in the aftermath of the crisis. Furthermore, this section comprises general information about the European Banking Union and a more detailed overview on the SSM. As a particular example, the differences between the SSM and former supervisory regulations in Austria are highlighted. In addition, the field of sentiment analysis is introduced in order to get an overview on this technology.

In a second step, potential data sources for the analyses are evaluated. This involves quantitative measures as well as text data, and is mainly done by reviewing related literature and interviewing banking experts. It is crucial to bear in mind that the sentiment analysis is done in the financial context. The concept of risk within this domain has to be clarified and one has to consider that words might be used in a way which is different to their common meaning. After clarifying these issues, the work deals with technologies for sentiment analysis. It provides a literature review and discusses methodologies and tools for the automatic extraction of risk sentiment from textual disclosures by banks.

The experiments and their evaluation are based on a corpus of suitable documents from a representative group of banks. Two independent approaches will be employed in order to find the best suited technique: first, a lexicon-based approach will analyze the evolution of risk sentiment. The second approach aims to predict quantitative risk indicators by means of supervised classification.

1.5 Structure of the Work

This master's thesis is divided in the chapters outlined below.

Chapter 2: Economic Background. This chapter sheds light on the financial crisis of 2007-08, which gave momentum to the discussions about a new European banking supervisory mechanism. It covers the chronology of events focusing on banks and disruptions in the financial markets. Furthermore, this chapter explains the European Banking Union's structure and compares the Single Supervisory Mechanism to the former supervisory regulations in Austria.

Chapter 3: Technical Background and Related Work. In this chapter, the thesis continues with an introduction to text mining and sentiment analysis. It presents relevant publications in the field and gives an overview on general approaches as well as important concepts of this technique.

Chapter 4: Data Sources and the Concept of Risk. The selection of proper data sources is crucial for successful risk sentiment analyses. This chapter evaluates different kinds of potentially suitable documents and discusses the concept of risk in banking. What are typical quantitative risk measures in banks and what does research reveal about risk culture in banks?

Chapter 5: Methodologies for Risk Sentiment Analysis This chapter describes the used approaches for risk sentiment analysis in the context of banking supervision. This involves the selection of a proper tool, data collection, and descriptions of both experiments' structure.

Chapter 6: Evaluation of the Experiments. The results of the experiments are used for evaluating the risk sentiment of a selected group of banks which will be supervised by the ECB. The extracted figures are the basis for analyses like the evolution of risk sentiment during the financial crisis and the predictability of quantitative risk indicators via machine learning techniques.

Chapter 7: Concluding Remarks. The last chapter concludes and highlights options for future research in this area.

CHAPTER 2

Economic Background

We are still dealing with consequences of the 2007-08 financial crisis, which caused a deep recession in many countries. The declining *Gross Domestic Product* (GDP) was accompanied by massive problems in the banking sector. These disruptive events fueled the discussions about a *European Banking Union* (EBU). This chapter gives a comprehensive overview on the financial crisis and the EBU. With regards to the focus of this thesis, particular attention is paid to the Single Supervisory Mechanism. This is relevant because the approaches developed in this thesis aim to improve risk assessments in the SSM by analyzing the sentiment of bank publications.

2.1 The Financial Crisis of 2007-08

The first part of this chapter covers the chronology of events and their impact on the economy. Furthermore, it summarizes some commonly agreed answers to the question *Why Did This Happen?* and describes the early responses to the crisis.

2.1.1 A Chronology of the Financial Crisis

From the 1980s on, Europe and many other countries experienced a long period of macroeconomic stability with low volatility of both GDP growth and inflation. Economists usually refer to this period as the *Great Moderation*. A key question is whether this period can be explained by structural changes or if it merely resulted from a longer period without severe macroeconomic shocks[1]. According to Cabanillas and Ruscher (2008), the answer lies in between. In their study about the Great Moderation in the Eurozone, they find by conducting regression analyses that "the Great Moderation is not just the result of a long period of luck in the form of milder shocks but can also partly be ascribed to changes in economic policies" (Cabanillas and Ruscher, 2008, p. 1).

[1]Macroeconomic shocks are events which influence aggregated demand and/or supply in an economy (cf. Blanchard and Illing, 2009, p. 242).

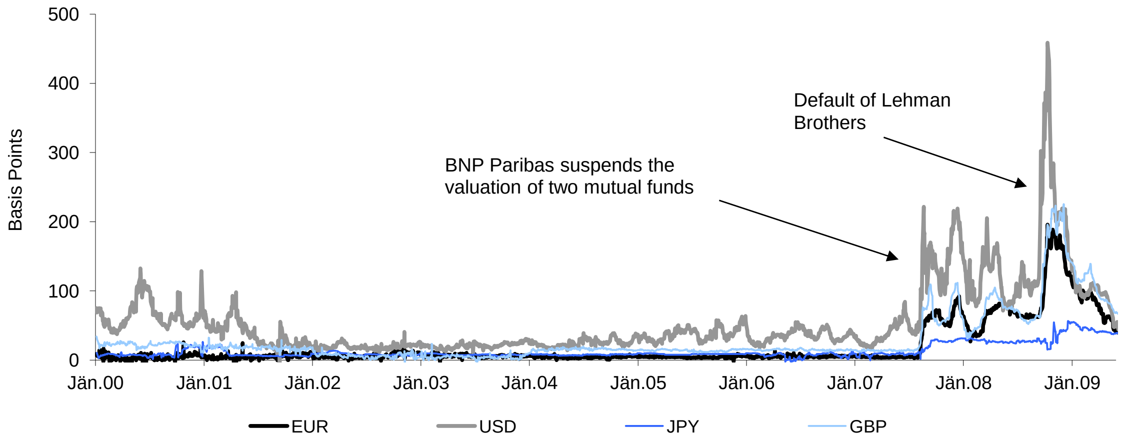

Figure 2.1: Difference between the interest rate on US government bonds and the 3-month interbank lending rate for several currencies, measured in basis points[3] (taken and slightly adapted from European Commission, 2009).

In Europe, we could observe that the shocks had different magnitudes across the countries, so the simple explanation of reduced *common* shocks does not hold[2]. Econometric analyses show some evidence that improvements in macroeconomic frameworks, particularly monetary policy, promoted reduced volatility in Europe (cf. Cabanillas and Ruscher, 2008, p. 3ff).

Economists at the European Commission concluded that the Great Moderation could be "a rather persistent feature of the euro-area economic landscape" (Cabanillas and Ruscher, 2008, p. 24), but it seems like this outlook was too optimistic. In fact, the crisis began already in the late summer of 2007, when banks started to face liquidity problems. When the French bank *BNP Paribas* stopped redeeming investment funds because of valuation problems, participants in the money market hesitated to lend to each other (Brunnermeier, 2009, p. 84). This resulted in rising interest rates in the interbank lending market. Figure 2.1 illustrates this mistrust which is measured by TED spreads, i.e. the difference between interest rates on US government bonds (T-bills) and the 3-month interbank lending rate (LIBOR). After several quiet years, the figures clearly indicate the problems in the banking sector from summer 2007 on.

A few months later, several banks failed and had to be rescued by governments. Notable examples of banks failed in spring 2008 are *Bear Stearns* in the United States or *Northern Rock*, a UK mortgage provider. Quaglia et al. (2009) point out that Northern Rock was ostensibly solvent, but potential lenders had doubts whether the bank was able to repay matured short-time loans. The resulting uncertainty was one reason for the (re)financing problems of Northern Rock.

[2] In the late 1980s and during the 1990s, most macroeconomic shocks were limited to individual countries. This was not the case in the 1970s (oil crises) and the 2000s (burst of dot-com bubble and financial crisis).

[3] Basis points are a common unit for measuring percentage changes in finance. A change of 1 % equals 100 basis points.

Further failures[4] led to concerns about huge systemic problems in the financial sector (cf. European Commission, 2009, p. 20f). The former *liquidity crisis* transformed to a *solvency crisis*. Governments were forced to rescue institutions, but financial integration in the EU sometimes complicated this task. An example is the Belgian bank *Fortis*, which faced liquidity problems in fall 2008. One of the reasons for this was that they took over the Dutch bank ABN-AMRO in 2007 as part of a consortium[5]. The governments of Belgium, Luxembourg and the Netherlands had to agree on a plan to rescue the bank, but the discussions ended in court (Quaglia et al., 2009, p. 78).

At about the same time, the three major banks in Iceland collapsed. After seven years of rapid growth in the banking sector, the government had to put *Landsbanki*, *Kaupthing*, and *Glitnir* into receivership. Because of advantageous interest rate spreads, these banks borrowed heavily abroad. Previously small local banks, they accumulated assets worth 11 times Iceland's GDP. When the value of these assets plummeted in the course of the financial crisis, the banks became insolvent. Even the Central Bank of Iceland was not able to fulfill its role as a *lender of last resort*, and the government had to take over the banks' domestic operations via an emergency law (Sigurjonsson, 2011). The events in Iceland are also a prime example for poor estimates of leading economists: in 2007, *Frederic Mishkin* of Columbia Business School published a report commissioned by the Icelandic Chamber of Commerce where he concluded that "[...] although Iceland's economy does have some imbalances that will eventually be reversed, financial fragility is currently not a problem, and the likelihood of a financial meltdown is low" (Mishkin and Herbertsson, 2006, p. 9).

On September 15th in 2008, the huge American investment bank *Lehman Brothers* went bankrupt. This unprecedented event showed that governments would not underwrite all risks. Financial markets reacted with a sharp rise of the spread between interbank lending rates and overnight swap rates. This figure is also referred to as the "fear index" (Quaglia et al., 2009, p. 66). According to European Central Bank (2010), the default[6] of Lehman Brothers showed some specific features making it disruptive:

- Since Lehman Brothers operated in several markets via subsidiaries and other interconnections, the default immediately affected the EU and other countries.
- The bank was deeply involved in exchange and derivative markets, served as issuer of financial instruments and dealt with structured products. Thus, Lehman was a counterparty in manifold business relationships all over the world.
- Due to their critical role in some markets, the bankruptcy negatively impacted the liquidity of financial institutions even when they did not directly interact with Lehman Brothers.

[4] European Commission (2009) names Lehman Brothers, Fannie Mae, Freddie Mac, AIG, Washington Mutual, Wachovia, Fortis, the banks of Iceland, Bradford & Bingley, Dexia, ABN-AMRO and Hypo Real Estate.

[5] Besides Fortis, the consortium consisted of the *Royal Bank of Scotland* and *Santander*.

[6] Since this thesis is written at the intersection of computer science, economics, and finance, a disambiguation of the term *default* is required: in the context of computer science, this term is used synonymously with *presets* or *standard settings* of a computer program. In finance and economics, it stands for the inability to make a payment (see also `http://www.merriam-webster.com/dictionary/default`, accessed October 10th, 2014).

These features illustrate the extensive cross-border consequences of the default. In fall 2008, the crisis worsened because of an adverse feedback loop (European Commission, 2009, p. 9): the economic downturn led to increased credit risks. Thus, banks had problems to raise capital which leads in turn to restricted lending. This vicious cycle is closed by the fact that restricted lending limits investments by companies, which amplifies the economic downturn discussed in the following section.

2.1.2 Impacts on the Economy

The crisis has its roots in the financial sector, but it soon spread over to the real economy via three main transmission channels explained in the first paragraph of this section. The other paragraphs outline the crisis' impacts on growth, employment, and fiscal costs.

Transmission Channels. The European Commission (2009) identified three main transmission channels explaining why the crisis spread from the financial sector to the real economy:

- Since credit losses reduced bank equity, the institutions had to reduce credit supply. Simulations suggest to view this channel as the primary problem with regards to its impact on economic activity. Especially in emerging markets, banks reduced funding via closures of credit lines and repatriation of capital.
- Increased household savings led to less demand for goods like cars and real estate. The increased savings can be explained with wealth and confidence effects.
- Not just consumer demand, but also business investment declined in fall 2008. This led to a global trade squeeze, which was intensified by the unavailability of trade finance and globalized trade chains.

Impact on GDP Growth. Figure 2.2 depicts the relatively constant inflation and growth of the GDP in the Eurozone until 2007. In 2008-09, the economy was hit by a hard recession. GDP in the Eurozone contracted by over 4 percent, making the crisis into the worst one since World War II. It is worth noting that countries in the European Union were affected to a different degree, depending on different initial conditions and vulnerabilities. Examples for diverging initial conditions are the extent of over-valuations in the real estate market, dependencies on exports, or the financial sector's relative importance (cf. European Commission, 2009, p. 27ff). In 2010, GDP in the Eurozone grew again with almost 2 percent. However, the IMF attests Europe in its *World Economic Outlook* just a "gradual and uneven recovery" (International Monetary Fund, 2010, p. 72). Especially the peripheral countries Greece, Ireland, Portugal, and Spain could not gain traction. According to the IMF, the EU was able to prevent the financial turmoil due to large-scale liquidity and credit support as well as substantial fiscal policy. Nevertheless, the Eurozone slipped back into a so-called *double dip* recession in 2012/13.

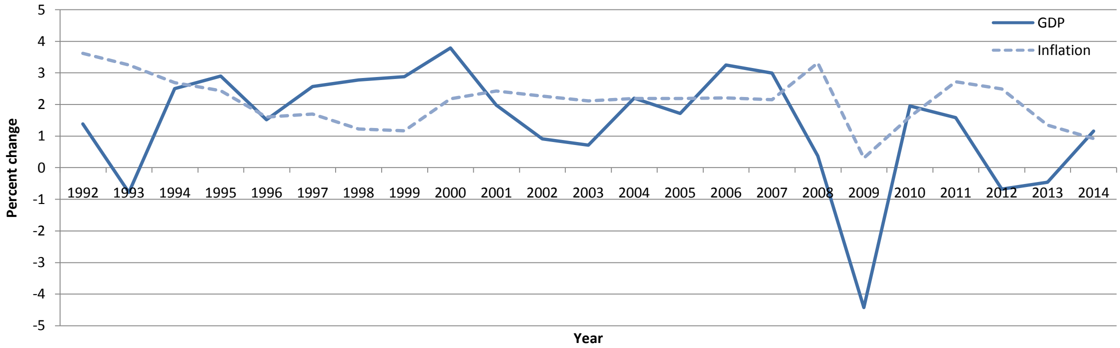

Figure 2.2: Percentual change of GDP and inflation in the Eurozone (own representation, data source: IMF World Economic Outlook Database, May 2014).

Impact on Employment. The recession was accompanied by serious problems in the labor markets. In 2009, 10 percent of the economically active population was unemployed (European Commission, 2010, p. 9). Especially the Baltic States, Ireland, and Spain faced high unemployment. Companies, trade unions and governments tried to keep employment relatively stable by utilizing flexible working time arrangements, and nominal wage concessions. This worked particularly well in Austria, Belgium, and Germany, where hours per worker were temporarily reduced. However, although such measures are very important, they can only help for a limited period of time since they do not improve the *structure* of labor markets (cf. European Commission, 2010, p. 9ff).

Impact on Fiscal Costs. The declining GDP reduced tax revenues, and the diverse measures for cushioning the crisis[7] increased government expenses. Naturally, this led to deteriorated public finances. From 2009 on, most European countries faced a fiscal deficit below the -3 % Maastricht criterion (cf. European Commission, 2009, p. 41ff). Since the majority of these deficits had to be financed by additional debt, interest payments were increasing as well. These developments led to a European sovereign debt crisis which is even threatening the common currency Euro. In this critical situation, it is a controversial issue whether governments should finance the deficits until economic recovery stabilizes or if it is better to use austerity measures in order to decrease debt. In Europe, liquidity-providing measures are bound to stringent fiscal consolidation. In of the most indebted European countries, Greece, consolidation measures did not turn out to be successful so far. In 2011, Greece's fiscal deficit was still 10 percent of GDP. However, aid programs from other countries *without* conditions are problematic as well since this would lead to moral hazard, i.e. Greece would probably not try hard enough to improve its situation if there are no incentives for reforms (cf. Collignon, 2012, p. 3ff).

[7]This includes both higher expenditures due to *automatic stabilizers* like unemployment benefits and discretionary stimuli like the *European Economic Recovery Plan*.

2.1.3 Why Did This Happen?

Economists, politicians and the comments in the media give different explanations for the financial crisis. The sheer complexity and diverse ideologies make it hard to give a clear diagnosis. This section summarizes the most agreed reasons for the events from 2007 on.

The financial crisis was triggered by the bursting US housing bubble and huge problems in the US subprime mortgage market. The term *subprime mortgage* is not clearly defined, but it is generally understood as risky mortgage where the borrower has a low creditworthiness because of previous delinquencies, foreclosures, bankruptcies, or a high debt service-to-income ratio (Sengupta and Emmons, 2007). Under normal circumstances, such borrowers would not be eligible for a real estate loan. Nevertheless, politicians in the US desired that every citizen should be able to buy his or her own home. This system worked as long as real estate prices were rising, because the homes acted as collateral—borrowers could even refinance their mortgages with the home's higher value.

For asset originators (e.g. the issuing bank), risks seemed to be low due to financial innovations based on the process of *securitization*. This process consists at least of the two steps depicted in Figure 2.3. The originator bundles the assets in a reference portfolio and sells it to another financial institution, the so-called *special purpose vehicle* (SPV). This first step ensures the desired off-balance-sheet treatment of the assets. In the second step, the SPV sells the pooled assets to investors. These asset-backed securities are usually tranched according to risk classes. *Senior tranches* are the safest ones because they will receive the portfolio's payments before others do. This is relevant in case of payment defaults when not all investors get their money back. The other extreme is called *junior tranche*. It is riskier, but generates a higher return than the senior tranches. Such tranches improve risk allocation since investors can decide how much risk they want to take (cf. Jobst, 2008, p. 48f).

Unfortunately, such products are complex and there is asymmetric information: the issuing bank has better knowledge about the individual risk than the buyer of asset-backed securities. However, due to securitization, such financial innovations received good ratings by rating agencies. Hence, buyers needed just a small amount of collateral as a guarantee[8]. When real estate prices went down, many subprime borrowers defaulted and had to leave houses or stop their construction. The number of defaults was so high that even the owners of senior tranches made losses. The lack of sufficient collateral made it impossible to stem the problems (cf. Aiginger, 2009, p. 2f).

While the US subprime crisis can be seen as a *trigger* of the financial crisis, there is also a number of deeper *root causes*. The following paragraphs give an overview on most agreed fundamental reasons.

Weakness in Coordination and Supervision. In a paper about the economic crisis, Karl Aiginger identifies room for improvement concerning European policy coordination: "It is necessary to coordinate European policy more closely internally as well as with those of the USA and with the dynamic economies of neighbouring countries and Asia in order to avoid further

[8] In this context, *collateral* denotes a proportion of the buyer's assets. If the borrower is not able to fulfill his or her obligations, the lender can seize these assets (cf. Bessis, 2002, p. 439).

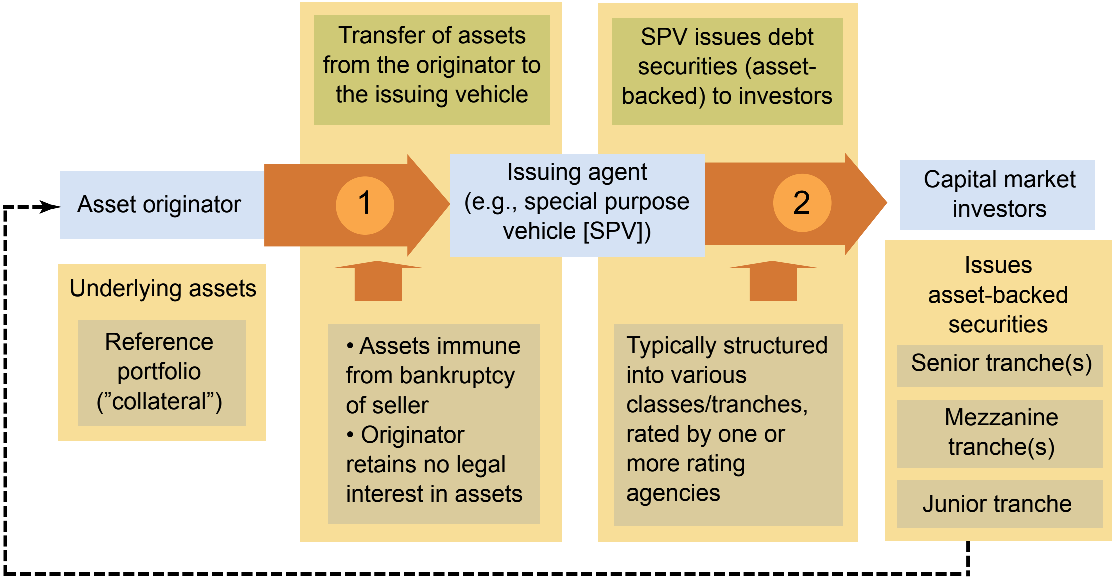

Figure 2.3: How securitization works (taken and slightly adapted from Jobst, 2008, p. 48).

crises, and proactively to tackle worldwide problems such as climate change, and raw materials and food shortage" (Aiginger, 2009, p. 1). Aiginger further criticizes rating agencies: coordinated stimulus programs are more effective than individual programs, but the agencies make such approaches less attractive since they rate individual governments and threaten them with downgrading if public expenditure is too high. Kapoor (2010) states that the supervisory structure lags behind the fundamental changes in financial markets. This affects not only Europe, but also the US and other countries. Consequently, Kapoor talks about the need for a *global* supervisory authority with statutory powers. National supervisors merely coordinate via informal meetings and Memorandums of Understanding, which is definitely not enough for dealing with global disruptions like the financial crisis. Kapoor sees the need for reforming national and international level supervision[9]: "The many gaps that existed in national and international level supervisory structures, together with the somewhat 'light touch' supervision in some countries, combined with poor co-ordination to give financial market actors carte blanche to engage in large scale regulatory arbitrage and build up excessive financial risks" (Kapoor, 2010, p. 55).

Misaligned Incentives and Inflated Expectations. Liberalization and deregulation in financial markets fostered economic growth in many countries. However, heterogeneous growth rates across countries and industries resulted in different returns. These differences and the profitable financial innovations put upward pressure on expectations concerning returns. This resulted in intensified speculation and risky behavior of investors (Aiginger, 2009). For Sony Kapoor, "misaligned incentives lie at the heart of financial crises" (Kapoor, 2010, p. 33). Remuneration

[9]Kapoor wrote the paper before the EU agreed on the banking union.

structures for bank employees promoted risky behavior since higher profits increased compensation. High profits[10] were realized through high leverages, for example through financial products derived from underlying securities. Taking high risks in the financial sector has systemic consequences. Because of the high degree of interconnectedness, a failing bank can drag down other financial institutions and causes costs across the whole sector. If the bank has to be bailed out by the government, the costs are passed to the taxpayers. On the other hand, "the benefits of excessive risk taking at a financial institution are reaped primarily by those who control it i.e. its employees and shareholders" (Kapoor, 2010, p. 37). Another example of misaligned incentives can be found in the compensation structure of subprime mortgage sellers: the more customers they acquired, the higher was their compensation. There was no incentive to do any background checks or to communicate any risks (Evans, 2010, p. 13).

Deregulation in the Financial System. Until the 1980s, banks faced extensive governmental restrictions like price limits and restraints of competition. Internationalization and growing importance of financial markets forced governments to relax some regulations. The goal was to increase efficiency of financial markets through competition. The first Basel capital adequacy rules were intended to prevent destabilizing effects of deregulations, but these rules did not fulfill expectations. Furthermore, existing regulations did not keep pace with the multitude of financial innovations like derivatives. The financial crisis unveiled systemic weaknesses in regulatory frameworks. They are too micro-oriented and static. Micro-orientation in this context stands for focusing on the stability of single institutes. Unfortunately, solvent banks do not guarantee systemic stability. Thus, economists suggest to improve macro-oriented regulation measures (cf. Hahn, 2008).

Monetary Policy in the US. After the burst of the *dot-com bubble* in 2000, the Federal Reserve lowered interest rates in order to promote growth. Critics argue that the Fed kept interest rates too low for a too long period of time, which fostered the housing boom in the US—even poor Americans could borrow money for their first home. It should be mentioned that the expansive monetary policy in the US also helped European export-oriented countries like Germany (Evans, 2010, p. 14f).

Global Financial Imbalances. Since the late 1970s, the US are running a current account deficit[11], nowadays primarily with China. These huge imbalances can be seen as as an "important causal factor behind the financial crisis of 2008 and the Great Recession" (Palley, 2011, p. 2), but politicians in the US did not see them as a big issue—the main goal in the 1980s was to keep inflation low. A popular explanation for the United States' current account deficit argues that the US Dollar's role as international reserve currency is the reason that many countries want to accumulate large quantities of US Dollars (Evans, 2010, p. 16f).

[10] Kapoor (2010) reports a bank return on equity (RoE) of 30 % in the UK just before the crisis hit.

[11] Hence, the United States import more goods and services than they export. Since they pay the imports with US Dollars, other countries accumulate this currency.

2.1.4 Early Responses to the Crisis

Governments, central banks and other institutions used several short-term measures in order to cushion the crisis, primarily monetary and fiscal policy, support for banks, guarantees, and coordinated actions by institutions of the European Union.

Central banks applied measures of *monetary policy* by reducing interest rates. In the US, the federal funds target rate is set to a range of zero to 0.25 % since December 2008. In Europe, the ECB followed a different policy: they lowered the main refinancing rate[12] throughout 2008 from 4.25 % to 2.50 % and down to 1.00 % in 2009. Since September 2014, this interest rate is at its record low of 0.05 %.

Governments utilized fiscal policy by introducing national programs for stabilizing the real economy. For example, Germany's fiscal stimulus amounted to 2.5 % of GDP over two years (Evans, 2010, p. 11). The goal of such measures is both to increase short-term demand directly and to stimulate long-term growth and employment (Aiginger, 2009, p. 12). Fiscal policy measures did not only consist of national programs, but included also *coordinated actions* by the EU, namely the *European Economic Recovery Plan* (EERP), adopted in December 2008. This plan combines coordinated fiscal policy and structural reforms. The long-run aim of this plan was to raise growth and job potential of the economy. In the short run, it "could support aggregate demand, employment and household income [...], whilst at the same time improving the adjustment capacity to enable a faster recovery when conditions improve" (European Commission, 2009, p. 71).

Another form of early response were *guarantees*. Their primary objective was to reduce uncertainty. Specific measures involved (temporary unlimited) guarantees for savings deposits, and the provision of guarantees on liabilities of distressed financial institutions. The latter were part of various measures which provided *support for banks*. In order to prevent the looming collapse of systemically important banks, governments set up rescue packages. The four main areas of support were debt guarantees, recapitalization, liquidity support, and treatment of impaired assets. The largest amount of money had to be used for state guarantees on bank liabilities (European Commission, 2009, p. 62).

2.2 Towards A European Banking Union

Soon after the financial crisis hit, calls for a system-wide banking supervision and regulation increased. The deeply integrated financial markets cannot be supervised via a bottom-up supervisory approach focusing on individual institutions: "That is why the European Union urgently needs to establish a supervisory body that has an eagle eye system-wide view of the financial system at least within the single market" (Kapoor, 2010, p. 42). Another author stated that the crisis "exposed the numerous flaws and fault lines of the pre-existing architecture of financial regulation and supervision" (Vourloumis, 2012, p. 1). As a matter of fact, the financial crisis also triggered official discussions about a comprehensive regulatory reform in the European Union.

[12] In particular, they lowered the fixed rate tender in the main refinancing operations; this and other key interest rates are published online at `http://www.ecb.europa.eu/stats/monetary/rates/html/index.en.html`, accessed May 14th, 2014.

An important document in this context is the *De Larosière Report* (de Larosière et al., 2009). It was mandated by the European Commission and carried out by a special high level group on financial supervision in the EU, chaired by Jacques de Larosière. The report was published in February 2009 and stressed the need for a new framework which involves the following items:

- A new regulatory agenda for dealing with risks, systemic shocks, pro-cyclical amplifiers, and incentives in financial markets.
- Stronger coordinated macroprudential and microprudential supervision[13] which builds on existing structures and preserves fair competition in the EU.
- Effective crisis management procedures for increased confidence and trust.

Based on the De Larosière Report, the EU started to build up the *European System of Financial Supervisors* (ESFS). This network of institutions consists of the *European Systemic Risk Board* (ESRB) and three *European Supervisory Authorities* (ESAs). The former's task is to detect sources of systemic risk threatening the financial system and to recommend strategies for dealing with such risks. The ESAs mediate between national supervision authorities and comprise the following institutions (Vourloumis, 2012, p. 12):

- *European Securities and Markets Authority* (ESMA): regulation of financial services, preservation of transparency and consumer protection.
- *European Banking Authority* (EBA): developing standards for banking supervision, protection of depositors and investors, preservation of financial stability.
- *European Insurance and Occupational Pensions Authority* (EIOPA): protection of policyholders, pension scheme members, and beneficiaries.

However, all these institutions kept supervisory competencies in the individual EU countries, and the European banking sector still faced problems. It became clear that the interconnected banking sector needs to be regulated and supervised directly at EU level within a *European Banking Union* (EBU), which completes the economic and monetary union. Its foundation is a "regulatory framework with common rules for banks in all 28 Member States, set out in a single rulebook" (European Commission, 2014). The EBU's objectives are to make banks more solid and immune to shocks. In case of a bank failure, the resolution should be financed by shareholders, creditors, and a resolution fund. The EBU will consist of three key elements explained in the next section.

[13]Microprudential supervision focuses on limiting the distress of individual financial institutions, whereas macroprudential supervision tries to limit problems in the overall financial system (de Larosière et al., 2009, p. 38).

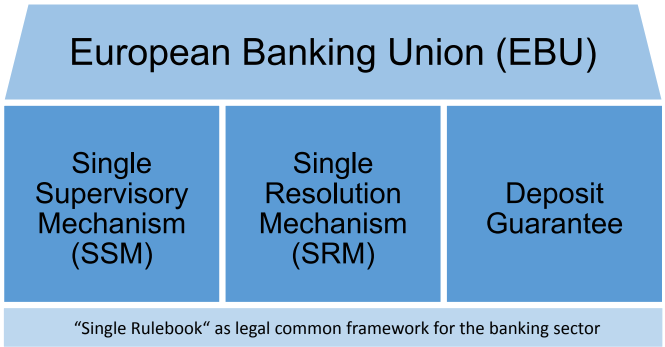

Figure 2.4: The European Banking Union (own representation based on Breuss, 2013).

2.2.1 The Three Pillars of the Banking Union

In June 2012, officials in the EU agreed on creating a Banking Union in the EU. The first step is the implementation of a *Single Supervisory Mechanism* (SSM). As Figure 2.4 shows, the Banking Union will be completed with the recently adopted *Single Resolution Mechanism* (SRM) and common rules for *deposit guarantees*. This section presents the most important facts about SRM and deposit guarantees, followed by another section with a more comprehensive analysis of the SSM.

The Single Resolution Mechanism. This mechanism is needed if banks fail despite improved supervision. It was adopted in April 2014 and provides rules for the resolution of cross-border banks. An important feature of the SRM is the protection of taxpayers: European Commission (2014) reports approved state aid for banks amounting to almost EUR 600 billion or 4.6 % of the EU's GDP in 2012. The SRM should avoid such a heavy burden for taxpayers. Instead, a "bail-in" mechanism takes shareholders and holders of bonds up on their promise—they would bear losses first.

The Single Resolution Mechanism's elements are the *Single Resolution Board* (SRB) and a *Single Resolution Fund* (SRF). The SRB is an agency consisting of four ECB representatives, the European Commission and the relevant national authorities. The SRF is funded by the banking sector and provides financial support for bank resolutions.

Figure 2.5 gives an overview on the process of resolution. This process involves ECB, SRB, the European Commission, and the Council of the European Union. The procedure of decision-making imposes a tight schedule in order to make accomplishment over one weekend possible: the ECB as supervisor notifies the SRB if a bank needs to be resolved. In cooperation with national authorities, the SRB defines the scheme used for resolving the bank and issues a recommendation. In the next step, the Council (Scenario 1) respectively the European Commission (Scenario 2) can object to the resolution scheme. The actual resolution is executed by the national authorities under supervision of the SRB (cf. Breuss, 2013, p. 18).

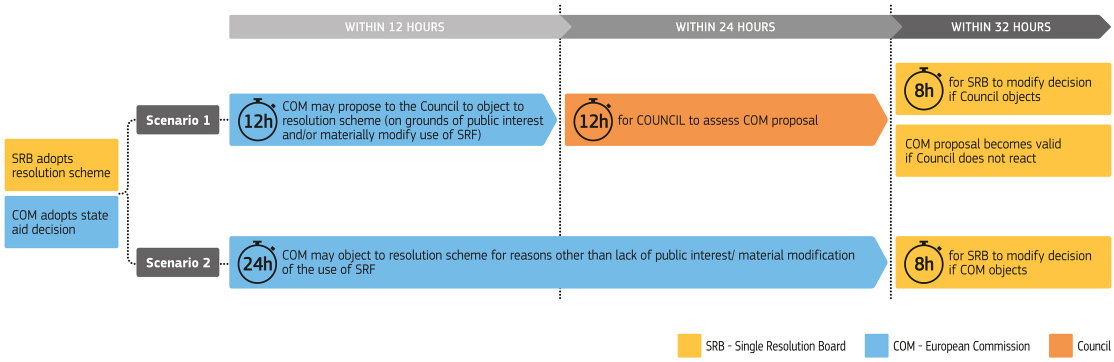

Figure 2.5: Resolution procedure in the European Banking Union (taken from European Commission, 2014).

Deposit Guarantee. A common and EU-wide deposit guarantee is the intended third pillar of the banking union, but it is not yet adopted—the establishment of SSM and SRM have been given a clear priority. According to Fritz Breuss, this pillar "is seen more skeptical and more or less rejected in the core countries of the euro zone (especially Germany), but it is advocated in the peripheral countries, as in the case of the rescue operations via the ESM" (Breuss, 2013, p. 25). Nevertheless, the EU already harmonized regulations after the financial crisis. The 2009/14/EC Directive ensures that deposits up to a minimum level of EUR 100,000 per depositor and bank are guaranteed even in the event of a bank failure. A draft legislation from 2010 proposes further steps of harmonization. In particular, an adoption of this draft would introduce common standards of financing, risk-based contributions, shorter payout period of seven working days, and improved handling of cross-border payout situations (International Monetary Fund, 2013b, p. 4).

Single Rulebook. The Single Rulebook denotes a EU-wide regulatory framework for banks and can be seen as the legal basis of the banking union. It is built up by 28 legal proposals, and banks within the Single Market have to comply with them (European Commission, 2014). The most important elements of this legal framework are the *Capital Requirements Directive and Regulation* (CRD IV / CRR), the *Bank Recovery and Resolution Directive* (BRRD), and the *Deposit Guarantee Scheme Directive* (DGSD). The Single Rule Book is completed with various other regulations for the EU's financial sector (Breuss, 2013, p. 12).

2.2.2 The Single Supervisory Mechanism

As a first step towards the European Banking Union, the European Commission published an initial proposal for a Single Supervisory Mechanism (SSM) in September 2012[14]. It gives the ECB the power to conduct prudential supervision of credit institutions, mainly in the Eurozone. The

[14] The proposal was published as COM(2012) 511 final.

final Regulation came into effect in November 2013. The Council of the European Union describes this mechanism in its SSM Regulation as "the system of financial supervision composed by the ECB and national competent authorities of participating Member States [...]" (Council of the EU, 2013, p. 63).

A *supervisory board* is responsible for planning and execution of the SSM's tasks. It consists of a Chair, a Vice Chair, four representatives of the ECB, and one representative of each participating state's NCA. The EU Council appointed *Danièle Nouy* as Chair and *Sabine Lautenschläger* as Vice Chair of the Supervisory board (European Central Bank, 2014d).

The SSM covers all countries whose currency is the Euro, but it is open for other states of the European Union as well. In the course of such a *close cooperation*, the member state has to comply with the ECB's guidelines and requests. Furthermore, it has to provide the necessary information for a comprehensive assessment of the member state's banks (Council of the EU, 2013, p. 77).

Significance of Banks. The European Central Bank directly supervises *significant* credit institutions via the SSM. In its SSM Framework Regulation, the ECB defines criteria for the assessment whether a bank is significant or not (European Central Bank, 2014b, p. 22ff):

- *Significance on the basis of size:* Generally, the size is determined by the total value of the bank's assets. If this sum exceeds EUR 30 billion, the institution is considered as significant.
- *Importance for the economy:* A bank is important and therefore significant for the economy of a member state if its total value of assets exceeds EUR 5 billion and if the national economic importance threshold is met, this is the case if the bank's total assets are greater or equal to 20 % of the state's GDP.
- *Cross-border activities:* If a supervised group's ratio of cross-border assets to its total assets exceeds 20 %, it is considered as significant. The same applies for cross-border liabilities.
- *Public financial assistance:* A bank is also significant when it requested or received direct public financial assistance from the European Stability Mechanism (ESM).
- *Three most significant banks:* If a supervised bank belongs to the three largest institutions in a particular state, it is considered as significant.

According to the ECB's *frequently asked questions* about banking supervision[15], 123 banks are considered as significant. While these institutions represent a rather small share of the *number* of banks[16], they hold with almost 85 % the majority of the total amount of assets. The non-significant banks are supervised by *National Competent Authorities* (NCAs)[17], but the ECB

[15] `https://www.bankingsupervision.europa.eu/about/faqs/html/index.en.html`, accessed January 29th, 2015.

[16] In total, there are about 6,000 banks in Europe (International Monetary Fund, 2013a, p. 15).

[17] For example, the *Financial market authority* (FMA) respectively the *Oesterreichische Nationalbank* (OeNB) in Austria or the *Federal financial supervisory authority* (BaFin) in Germany.

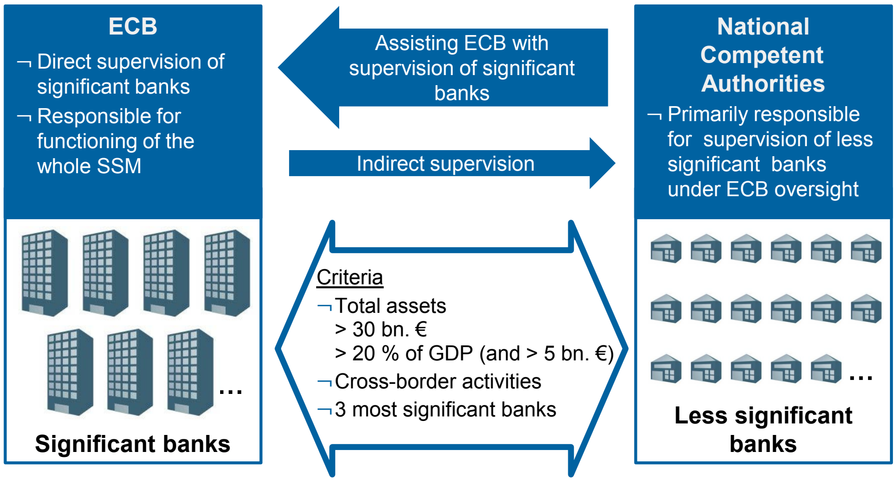

Figure 2.6: SSM structure and cooperation (taken and slightly adapted from Sprenger, 2013).

is "responsible for the effective and consistent functioning of the SSM" (Council of the EU, 2013, p. 75). Figure 2.6 shows the SSM's basic structure and illustrates the cooperation between ECB and the NCAs.

Comprehensive Assessment. One of the ECB's first preparation steps for the SSM was to conduct a *comprehensive assessment* of the significant banks. The objective was to increase confidence in the financial sector by getting better information about the condition of banks, and by undertaking corrective actions in balance sheets. Launched in October 2013, the comprehensive assessment consisted of two elements. The first one was a broad and inclusive *asset quality review*. The objectives were to get an overview on bank exposures and to assess the quality of their assets. The reviews were carried out by the NCAs based on centrally specified requirements and methodologies. The second element was a *stress test*. In the course of execution, the ECB worked closely with the EBA and incorporated the asset quality review's outcome (cf. European Central Bank, 2014c, p. 13ff).

Both tasks were finished by the end of October 2014. Furthermore, a *supervisory risk assessment* supplemented the comprehensive assessment. Its objective was to assess "a bank's intrinsic risk profile, its position in relation to peers and its vulnerability to a number of exogenous factors" (European Central Bank, 2014c, p. 15). The risk assessment further acted as consistency check for the results of the comprehensive assessment.

Cross-Country Experiences. In a paper about the European Banking Union, the International Monetary Fund (2013a) draws some comparisons to other federations, in particular the United States. In the US, several agencies share responsibility for supervision, resolution, and safety nets. Central banking and supervision in the US is organized within the Federal Reserve System. The central *Board of Governors* delegates banking supervision to the twelve regional reserve banks. In contrast, the authority responsible for deposit insurance works fully centralized. A central supervisory structure in Europe, however, is challenging because of legal and cultural issues. At the national level, there are different laws, accounting and tax frameworks, and languages. The IMF concludes that "design of this interaction between the center and national authorities, of the decentralization and delegation of tasks, and the allocation of powers between the two levels will play a key role in achieving effective supervision" (International Monetary Fund, 2013a, p. 13). Other crucial factors for successful single supervision in Europe are adequate resources at the ECB, strong governance, and a complete framework. According to the IMF, the latter also requires a common safety net, which is not yet adopted in the European Banking Union.

Risk Assessment in the SSM. The detection of risk is of particular relevance for this work. The first SSM Quarterly Report (European Central Bank, 2014c) provides some information about the *Risk Assessment System* (RAS). As already mentioned in the introduction to this work, the RAS should utilize both quantitative and qualitative indicators for investigating a bank's intrinsic risk profile. In addition, the RAS aims to evaluate the bank's position relative to its peers and its vulnerability to exogenous threats.

The categories for assessing the risk are as follows: "business risk and profitability, credit and counterparty risk including residual risk, interest rate risk in the banking book, market risk, securitisation risk, operational risk, insurance risk, concentration risk (inter-risk), internal governance and risk management, liquidity risk and capital position" (European Central Bank, 2014c, p. 10). After assessing these individual categories, the results are combined to an overall assessment.

Data Reporting and Technical Standards. The SSM Framework Regulation confers the process of collecting and reviewing data reported by banks to the ECB (European Central Bank, 2014b, p. 48). This supervisory data reporting framework is also needed for the Risk Assessment System described above. A specific database within the framework is the *Supervisory Banking* (SUBA) data system. It will be used for handling data and meta data reported by individual banks and groups. Another database is the *Register of Institutions and Affiliates Database* (RIAD), which will be linked to SUBA (European Central Bank, 2014d, p. 16f).

The supervisory data reporting framework has to comply with relevant EU law and the *Implementing Technical Standards* (ITS). These standards have been developed by the European Banking Authority (EBA) on basis of *Common Reporting* (COREP) and *Financial Reporting* (FINREP) guidelines. The ITS are part of the single rulebook and specify "uniform formats, frequencies, dates of reporting, definitions and the IT solutions to be applied by credit institutions and investment firms in Europe" (European Banking Authority, 2013, p. 3). Generally, harmonized reporting frameworks are needed for consistent risk assessments across Europe and for reducing information asymmetries between supervisory authorities. This improves the internal market's efficiency and reduces regulatory arbitrage opportunities. The main features of the ITS are (cf. European Banking Authority, 2013, p. 3f):

- *Financial information:* the rationale of including financial information in the ITS is to gain information on the risk profile of individual institutions and to assess systemic risks. Furthermore, Article 99 of the *Capital Requirements Regulation* (CRR) requires institutions which apply *International Financial Reporting Standards* (IFRS) to report such financial information.

- *Proportionality, frequency and reporting dates:* the principle of proportionality exempts some institutions[18] from reporting burdensome data points. Uniform reporting frequencies and remittance dates are set according to the type of information and the resulting administrative burden.

- *IT solutions for data submission:* the EBA develops taxonomies for the XML-based *eXtensible Business Reporting Language* (XBRL) based on a formal data point model. NCAs and banks are encouraged to use this taxonomy for ITS implementations.

- *Time frame for ITS implementation:* although institutions have to fulfill the new CRR requirements since January 2014, some remittance periods have been extended in order to provide more time for implementing ITS requirements. For example, ITS requirements relating to financial information had to be met only from the third quarter of 2014 on.

2.2.3 Banking Supervision in Austria

In order to give an example about the SSM's influence on the National Competent Authorities (NCAs), this section summarizes developments regarding banking supervision in Austria.

Originally, the Federal Ministry of Finance was responsible for supervising banks in Austria. The *Oesterreichische Nationalbank* (OeNB), Austria's central bank, performed examinations of banks. A 2001 report on reforming financial supervision in Austria identifies several weaknesses like the lack of clear division of responsibilities between the Ministry of Finance and the OeNB, or the absence of clear objectives (Zechner et al., 2001).

These weaknesses led to fundamental reforms. In April 2002, the new *Financial Markets Authority* (FMA) started operations. The FMA was founded as an independent supervisory authority and improved Austria's supervisory structure. However, this step did not eliminate some systemic weaknesses, which arose primarily because of overlapping responsibilities between OeNB and FMA. Another reform in 2008 aimed to increase efficiency and effectiveness of banking supervision in Austria. Specific measures include closer monitoring of institutions, a central platform for the electronic submission of reports, and better decision-making structures. The OeNB conducts on-site bank inspections as well as off-site analyses of banks and financial conglomerates. In addition, the OeNB is responsible for regular bank inspections which serve as basis for the FMA's official actions. The FMA is the authority in charge of banking supervision. Furthermore, it monitors trading of securities and handles licensing issues (cf. Oesterreichische Nationalbank and Austrian Financial Market Authority, 2009).

Restructuring Banking Reporting in Austria. In the 1990s, the OeNB designed tailor-made reporting forms for the collection of various data. Unfortunately, this approach lacked consistency and made it hard to maintain an overview of the data collection process. Despite the

[18] In particular institutions with insignificant activities or negligible systemic importance.

institutional reforms in the last decade, reporting forms in Austria remained complicated and inefficient. Additionally, the number and complexity of required reports increased. These developments made it necessary to restructure Austria's banking reporting system (Hille, 2013, p. 1f).

The new reporting system was developed together with Austrian banks. First discussions took place in 2011, and working groups have been conducting specific preparations since 2013. At the new system's core is the *basic cube*, an abstract data model which defines harmonized reporting standards at a very granular level. Specifically, it defines individual business transactions and their attributes and makes the data comparable across banks.

Multidimensional *smart cubes* offer specific aggregation and selection algorithms and the basic cube's data for generating reports. These reports are used for satisfying national and international obligations. Because of practical issues and legal constraints, the OeNB developed two types of smart cubes: the first one discloses details to the central bank (micro data), the second one does not (macro data).

Since the basic cube is an abstract data model, the cooperating banks have to implement it within their database structure. An *Entity Relationship* (ER) model[19] serves as representation of the basic cube. Examples for entities in this model are *contract*[20], *counterparty* or *collateral*. For the definition of the attributes, it was crucial to build up a common dictionary for reporting terms. This seemed to be challenging: "Our experience has shown us that this is a rocky road because the representatives of the banks need to improve their understanding of the concepts behind the reports. For their part, the representatives of the central banks have to learn how to be more specific in defining what they require from the banks and how they best can formulate these definitions more precisely" (Hille, 2013, p. 273).

Implications of the SSM on Austrian Banking Supervision. Interestingly, the paper about the new Austrian reporting model hardly mentions issues related to the European Banking Union. However, Hille is convinced that the model fits all needs. He states that "the introduction of new reporting requirements is expected to be very easy if the attributes necessary to produce the smart cubes are available in the basic cube" and that "[...] there is a plan to cover the DPM defined by the EBA"[21] (Hille, 2013, p. 281).

Huber and von Pföstl (2013) emphasize the new responsibilities of OeNB and FMA within the SSM. As mentioned above, one representative of each NCA will be a member of the SSM Supervisory Board. Participation in this board implies involvement in decision-making at the European level, i.e. decisions regarding the largest banks in the Eurozone. Although significant banks are supervised directly by the ECB, experts from NCAs are involved as well since they are responsible for on-site examinations.

According to Austrian Financial Market Authority (2014), preparations for the SSM showed that six Austrian banking groups are going to be considered as significant. With *Unicredit Bank*

[19]ER models are a standard approach for data modeling. They consist of ER diagrams and descriptions of the used elements. ER diagrams are built up by entities, their attributes and relationships between the entities.

[20]For example deposits, savings accounts, or investments in securities.

[21]The *Data Point Model* (DPM) mentioned in this quote is a structured representation of the data the *European Banking Authority* (EBA) wants to collect.

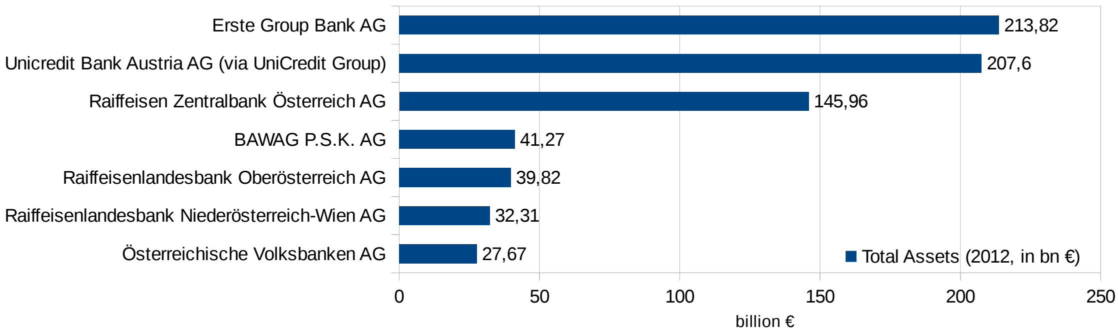

Figure 2.7: Significant banks in Austria (own representation based on Austrian Financial Market Authority, 2014, p. 30, and bank websites).

Austria AG, a seventh Austrian bank is supervised by the ECB via the Italian *UniCredit Group*. Figure 2.7 lists these groups including their total assets. The remaining 650 Austrian banks are smaller and continue to stay under supervision of the FMA. Since many Austrian banks have a high exposure to Eastern Europe, the FMA would welcome participation of non-euro countries in the SSM.

2.3 Conclusions

After a long period of macroeconomic stability which started in the 1980s, a severe financial crisis struck the economy. The crisis was initially limited to the financial sector, but it quickly spread over to the real economy and led a contraction of the GDP in the Eurozone by over 4 percent in 2008-09. Short-term responses to the crisis included support for banks as well as guarantees, but this did not cure the underlying problems.

In order to promote long-term stability in the financial sector, the EU decided to supervise the largest banks directly via the Single Supervisory Mechanism, which is a part of the European Banking Union. Risk assessments in the SSM are based on harmonized reporting frameworks and should include both quantitative and qualitative indicators. The objective of this thesis is to explore how such risk assessments could be supplemented by analyzing textual data published by banks. In particular, this is done by employing *sentiment analysis*. The next chapter will introduce this popular approach in the field of text mining.

CHAPTER 3

Technical Background and Related Work

As stated in the previous chapters, banking supervision within the Single Supervisory Mechanism also comprises qualitative analyses. *Text mining* approaches have the potential to play a meaningful role in this process. In particular, *sentiment analysis* could be used for improving risk assessments of banks by computational analyses of company disclosures and other relevant documents. This chapter covers the technical background knowledge by introducing text mining and sentiment analysis. Additionally, it presents related work in order to give an overview on the *state of the art*.

3.1 Text Mining

The evolution of the web, social networks and other online media promoted the creation and availability of textual data in vast amounts. In contrast to structured data, text documents are not managed via databases. They are typically accessed by using search engines. The corresponding task is *information retrieval*. The field of *text mining* has a different focus. Its primary goal is to discover *new patterns* by analyzing textual information and it goes "beyond information access to further help users analyze and digest information and facilitate decision making" (Aggarwal and Zhai, 2012c, p. 2). Another characterization of the field states that text mining is about "discovery and extraction of interesting, non-trivial knowledge from free or unstructured text" (Kao and Poteet, 2007, p. 1).

Text mining makes use of *natural language processing* (NLP) techniques for extracting data from texts, for example *stemming* or the detection of negations[1]. NLP deals with parsing and understanding natural language and "makes use of linguistic concepts such as part-of-speech (noun, verb, adjective, etc.) and grammatical structure" (Kao and Poteet, 2007, p. 1). *Statistical*

[1]Both techniques are explained in Section 3.3.1.

NLP is of particular relevance for this thesis, it comprises "all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra" (Manning and Schütze, 1999, p. xxxi).

In their *Introduction to Text Mining*, Aggarwal and Zhai (2012c, p. 4ff) present algorithms and applications for text mining. The following paragraphs present some of them as an introduction to the field.

Information Extraction from Texts. This general approach aims to extract semantics from text data instead of merely counting specific words. Aggarwal and Zhai (2012c) give the extraction of entities and their relations as an example. These data could be used for revealing hidden structures in the text by conducting network analyses.

Text Summarization. One way for giving an overview on an extensive document is to pick elements from the original text. This method is called *extractive summarization*. Another way is to conduct *abstractive summarization*. In this case, the algorithm constructs a summary which does not necessarily contain sentences from the original document.

Unsupervised Learning Methods. These methods work without training data and can therefore be applied to texts without manually labeling them. *Clustering* assigns each document to a partition representing a topical cluster. However, usually it is hard to assign a document unambiguously to a single cluster. *Topic Modeling* computes probabilities for cluster memberships of documents. Hence, it is a *soft* clustering method and each *topic* has a corresponding probability. In a next step, documents can be represented "as a linear probabilistic combination of these different topics" (Aggarwal and Zhai, 2012c, p. 5).

Supervised Learning Methods. Many data mining applications deal with *classification*. Such tasks are often tackled with supervised learning methods. By utilizing training data, these methods construct a classifier for computing predictions on new text data. Examples for specific applications are "target marketing, medical diagnosis, news group filtering, and document organization" (Aggarwal and Zhai, 2012a, p. 163).

Cross-Lingual Mining of Text Data. There are applications of text mining where the document corpus includes text data written in more than one language. Nie et al. (2012) survey the use of cross-lingual text mining in two fields, namely *cross-lingual information retrieval* (CLIR) and *machine translation* (MT).

The objective of CLIR is to retrieve documents in a language which differs from the original query's language, whereas machine translation aims to transfer text data from one language to another. Classical MT systems are rule-based and work with dictionaries. More recent approaches employ large corpora of bilingual text data with the objective of obtaining *translational knowledge*, which is "in its essence, a form of translingual text mining" (Nie et al., 2012, p. 324).

Text Mining in Social Media. The massive increase in social media activity generated large amounts of text data containing subjective information. Hence, these data are interesting for marketers since they allow for gaining better customer insights. Challenges of text mining in social media are the dynamic context, and the use of slang and poor vocabulary. In social network mining, it is also important to analyze links rather than just the textual content.

Opinion Mining from Text Data. Opinionated texts like product reviews comprise valuable information. There are many applications for text mining in this context, for example decision-support systems for consumers, business intelligence, and the detection of "spam opinions" with low usefulness.

3.2 General Introduction to Sentiment Analysis

In their *Survey of Opinion Mining and Sentiment Analysis*, Liu and Zhang start with a general definition of this topic: "Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes" (Liu and Zhang, 2012, p. 415). Pang and Lee (2008) state that many papers dealing with sentiment analysis focus on analyzing the polarity of reviews, i.e. they are viewing it as a classification task. Nevertheless, they suggest defining the field in a broader sense and denote it as the "computational treatment of opinion, sentiment, and subjectivity in text" (Pang and Lee, 2008, p. 6).

After a general introduction, this section clarifies the terminology, presents selected applications of sentiment analysis and closes with some remarks on challenges and limitations.

3.2.1 General Remarks

In the 1960s, the *General Inquirer* was introduced as one of the first computer approaches to content analysis (cf. Stone et al., 1966). The authors describe the rationale and procedures of the system, and define content analysis as a "technique that enables inference by objectively and systematically identifying specified characteristics within text" (Kearney and Liu, 2014, p. 2). Other early papers dealt with the computational detection of beliefs, metaphors, or affect (Pang and Lee, 2008, p. 4). For example, Carbonell (1979) wrote an early PhD thesis on computer models of belief systems.

These examples show that researchers have been interested in analyzing textual data for a long time, but only the broad emergence of the web led to intensified efforts in this field of research. Social media and other tools fostered the creation of large amounts of opinionated text available digitally. Within the last decade, "sentiment analysis has grown to be one of the most active research areas in natural language processing" (Liu, 2012, p. 5). It evolved to an interdisciplinary field used not just in computer science, but also in management sciences and social sciences. Commercially oriented events like the *Sentiment Analysis Symposium*[2] demonstrate a high practical relevance of this topic with applications in various domains (cf. Liu, 2012). Pang and Lee (2008) name similar reasons for the interest in this subject:

[2] See `http://sentimentsymposium.com/`, accessed June 5th, 2014.

- The rising popularity of machine learning methods in natural language processing and information retrieval,
- the availability of training data sets for supervised learning algorithms,
- and the realization of manifold commercial and intelligence applications.

A Note on Subjectivity. An area of research within sentiment analysis is *subjectivity detection*. This task aims to "decide whether a given document contains subjective information or not" (Pang and Lee, 2008, p. 18). Deciding on subjectivity is challenging since the difference is often subtle. Take the sentence "the battery lasts 2 hours" as an example. One cannot conclude from this sentence whether the author is satisfied with the 2 hours or not. On the other hand, the statement "the battery only lasts 2 hours" also contains subjective information. The word *only* indicates that the author expected a longer battery life. However, the aim of this thesis does not require strictly opinionated texts since even *objective* information can help to assess the overall sentiment (cf. Pang and Lee, 2008, p. 16ff). For example, the following sentence has a positive connotation in a business context although it does not contain opinionated words: "In December 2013, we re-launched the BAWAG P.S.K. website, aiming at providing our customers with easier, more efficient and more secure access to banking products and services" (BAWAG P.S.K., 2014, p. 43).

Domain Considerations. Sentiment analysis is usually domain dependent. The same words and phrases can have completely different meanings—even in related contexts. Consider the advice "go read the book". This is an unambiguously positive statement in a book review. In a movie review, however, it would be a negative sentiment. Several algorithms try to solve the issue of domain transfer, e.g. the *structural correspondence learning algorithm*. The latter depends on pivot features in both domains which are used for constructing a feature mapping matrix.

Another approach is to construct domain-specific dictionaries or annotation databases. Within the context of this thesis, it is worth to cite Loughran and McDonald (2011) and Takala et al. (2014). These authors constructed specific word lists respectively annotations for texts in the financial and economic context.

3.2.2 Terminology

There are two common terms which denote, at least in essence, the same field: *opinion mining* and *sentiment analysis*. However, the common usage of the words opinion and sentiment differs. According to the *Collins Dictionary*, opinion is about "a belief not based on absolute certainty or positive knowledge but on what seems true [...]"[3], whereas sentiment is "a complex combination of feelings and opinions [...]"[4], often based on careful consideration. Considering these differences, the term *sentiment analysis* seems to be more appropriate for this thesis.

[3] See `http://www.collinsdictionary.com/dictionary/american/opinion`, accessed June 6th, 2014.

[4] See `http://www.collinsdictionary.com/dictionary/american/sentiment`, accessed June 6th, 2014.

Pang and Lee (2008) note that the term *opinion mining* is particularly popular among researchers dealing with web search or information retrieval, and is used for analyzing various kinds of evaluative texts. The term *sentiment analysis* has been around since the early 2000s, when studies on market sentiment were published. Nowadays, the term is popular in the field of natural language processing. In their *Survey of Opinion Mining and Sentiment Analysis*, Liu and Zhang (2012) give the following formal definition of an opinion:

A *regular opinion* is a quintuple $(e_i, a_{ij}, oo_{ijk}, h_k, t_l)$ consisting of the following elements:

- Name of an entity e_i, which could be a person, product, topic, or organization.
- Aspects of the entity a_{ij}, *GENERAL* is used if the opinion is on the entity as a whole.
- Orientation of the opinion about an aspect of the entity oo_{ijk}, for example *positive / neutral / negative* or *strong / neutral / weak*.
- Holder of the opinion h_k.
- Time when the opinion is expressed by its holder t_l.

It is useful to illustrate this definition with a piece of opinionated text related to this thesis. The following paragraph is taken from the *Letter from the Chief Executive Officer* in the *BAWAG P.S.K. Consolidated Annual Report 2013* and annotated with numbers:

> "(1) In 2014 and beyond, we as well as other banks in Austria and Europe will continue to face headwinds. (2) Specifically, we plan for a persistently low interest rate environment, increased competition and regulatory challenges, including new requirements from the ECB under its Single Supervisory Mechanism. (3) Throughout this difficult environment, we will continue to strongly position BAWAG P.S.K. as the most efficient bank on the Austrian market" (BAWAG P.S.K., 2014, p. 11).

These statements have several interesting features. The *entity* is always BAWAG P.S.K. itself. Sentence (1) is a general statement about the bank's future implying a challenging business environment. For an automated sentiment analysis system, it is most likely challenging to interpret "headwind" in the correct way. In sentence (2), the phrases "increased competition" and "regulatory challenges" also have to be considered as negative, while the other two pieces of information are neutral. Sentence (3) contains a *comparative opinion* expressed by the superlative "most efficient bank on the Austrian market". An example for an opinion quintuple generated from sentence (1) would be (BAWAG P.S.K., GENERAL, negative, CEO, 2013). The selected paragraph does not reveal any information about the CEO's attitude towards risk, but one can derive that the bank manager anticipated an uncertain and therefore risky business environment in 2014.

3.2.3 Applications

In addition to the popular analyses of reviews, there are many other applications of sentiment analysis. People's opinions and attitudes are important for many stakeholders—both in personal life and for companies. Pang and Lee (2008) group applications in the following broad categories:

Business and Government Intelligence. As Pang and Lee point out, "business intelligence seems to be one of the main factors behind corporate interest in the field" (Pang and Lee, 2008, p. 8). Sentiment analysis can help to gain insights beyond objective advantages and disadvantages of product features, e.g. whether customers like the design or not. Possible government intelligence applications involve tracking public viewpoints or shifts in hostile communications (Pang and Lee, 2008, p. 8f). However, all these applications potentially face privacy issues.

Besides websites dedicated to product reviews, social media are an important source of information for business intelligence applications. Posts on *Twitter* and *Facebook* are primarily used for monitoring brand reputation (Feldman, 2013).

The domain of financial markets is another area where sentiment analysis can play to its strengths. A possible application is the automated aggregation of opinionated documents about a firm into a single score. Feldman (2013) names *The Stock Sonar*[5] as an example. This system aggregates positive and negative sentiment and displays these data alongside the stock price. The domain of this thesis, risk sentiment analysis, can also be considered as an application of business respectively government intelligence.

Review-Related Websites. The web offers a multitude of review-related sites. They contain opinions about products, hotels, movies and so forth. Finding and monitoring relevant sites and summarizing reviews is a cumbersome task due to the huge volume of text. Automated sentiment analyzers can help to resolve this issue (Liu, 2012, p. 8f). Other specific applications in this area are the summarization of user reviews and fixing erroneous ratings. The latter is needed when users accidentally chose a low rating although the corresponding text was positive. Furthermore, reviews could be biased (Pang and Lee, 2008, p. 7).

Sub-Component Technologies. As a sub-component, sentiment analyzers are an enabling technology for other systems. For example, using them within recommendation systems could improve the latter by utilizing textual information in reviews. Another application is the improvement of ad placements on websites, e.g. placing ads only when positive sentiments about this product are detected. A third system where sentiment analysis could be used as a sub-component are question answering systems, especially when opinion-related questions are asked. Finally, it would be beneficial for citation analyses if one could determine whether a citation is used as supporting evidence or as negative example (Pang and Lee, 2008, p. 7f).

Applications in Politics, Sociology and other Domains. In politics, sentiment analysis can be used for exploring the attitudes of voters or politicians. Furthermore, there are promising

[5]See `http://www.thestocksonar.com`, accessed June 11th, 2014.

applications within sociology. Sentiment analysis can help to assess trust and attitudes among persons, which affects, for example, group cohesion. Since opinions and sentiment are a relevant factor in various domains, Pang and Lee (2008) assume that sentiment analysis is an interesting technology for many other fields as well.

3.2.4 Challenges and Limitations

Sentiment analysis is a difficult task, especially because of the fact that sentiment in written statements is often ambiguous. If two humans independently rate the sentiment of documents, they will not agree on every rating. For automated approaches, a proper selection of opinion words is crucial, but still not a guarantee for successful sentiment analyses. Feldman (2013) and Liu (2012) discuss several issues which arise because of the flexibility of natural language and other factors:

- As already discussed, the *domain dependence* of opinion words is a major challenge.

- Sentiment words do not necessarily reveal an opinion. In particular, this is often the case in questions and conditional sentences. Consider the following conditional sentence from an annual report: "If the risk potential exceeds a defined limit, the implementation of suitable measures is mandatory" (BAWAG P.S.K., 2014, p. 183). In fact, this sentence does not imply a high risk potential although it contains the words *risk* and *exceed*.

- In contrast to the previous item, sentiment can also be expressed *without* using opinion words. This can be the case when factual sentences imply an opinion like in the statement already cited above: "In 2014 and beyond, we as well as other banks in Austria and Europe will continue to face headwinds" (BAWAG P.S.K., 2014, p. 11). This is actually objective, but implies a negative outlook for 2014. Most sentiment analysis systems focus on subjective sentences and therefore overlook such information.

- The issue of *automatic entity resolution* is still challenging: algorithms should detect if texts use more than one name for the same entity. A related issue is grouping terms which refer to the same aspect, e.g. *battery life* and *power usage*.

- In text with multiple entities, sentiment analysis systems have to match text elements with the respective entities. However, "current accuracy in identifying the relevant text is far from satisfactory" (Feldman, 2013, p. 88).

- It is hard to detect *sarcasm* automatically, e.g. in "What a great car! It stopped working in two days" (Liu, 2012, p. 13). Approaches for automated sarcasm detection are still in their infancy.

- Other challenges are the use of slang words, missing or wrong punctuation, misspellings, grammatical mistakes, also referred to as *noisy texts*. In particular, this is an issue for analysis of text in social media and web forums.

With regards to this thesis, several of these items restrict the validity of the results, while others are less problematic. For example, it is sufficient if the selected approach is able to deal with the financial domain. Moreover, the analyzed documents discuss just a single entity, i.e. an individual bank. It is further unlikely that documents like annual reports or newspaper articles use slang words or contain spelling mistakes. Nevertheless, it is important that the selected approach is able to extract implicit sentiment from objective sentences. Another challenge is the correct interpretation of conditional sentences.

3.3 Important Concepts and Approaches for Sentiment Analysis

This section starts with introducing the concept of *features* as well as methods for term weighting and feature selection. Afterwards, the section continues with an introduction of the "two main approaches to the problem of extracting sentiment automatically" (Taboada et al., 2011, p. 268), namely *lexicon-based approaches* and *supervised learning*.

3.3.1 Features

Sentiment analysis operates with *features* of the text, which are essentially attributes of the analyzed data. A simple example for a feature is an individual word, also referred to as *unigram* (Liu and Zhang, 2012, p. 423). An effective set of features is crucial for sentiment analysis, so there is a huge body of research dealing with this topic. The following paragraphs discuss some popular features used in sentiment analysis (cf. Pang and Lee, 2008; Liu and Zhang, 2012).

Opinion Words and Phrases. Obviously, opinionated words are important features since they express certain sentiments. The literature denotes opinion words also as *opinion-bearing words* or *sentiment words*. Table 3.1 lists examples for opinion words in the financial context.

Positive	efficient, great, improved, stabilized, success, vibrant
Negative	abrupt, closure, drastically, fail, postpone, threat
Uncertainty	approximately, doubt, might, probably, risky, volatility

Table 3.1: Examples for opinion words in the financial context (cf. Loughran and McDonald, 2011).

As Table 3.1 shows, not only adjectives and adverbs can be used as opinion words. Many nouns and verbs can be associated with sentiments as well, for example *annoyance*, *bankruptcy*, *disaster*, *hate* and *like*. Besides single words, phrases and idioms can also bear sentiment. An example from the business context is *to gain ground*, a phrase which is used synonymously with *making progress*.

Parts-of-Speech (POS). These are linguistic features occurring in natural languages. They basically describe word classes like nouns, adjectives, articles, or conjunctions (Voutilainen, 2003). Parts-of-speech can be considered as a "crude form of word sense disambiguation" (Pang and Lee, 2008, p. 21), making it possible to use POS as features for sentiment analysis. As indicated in the paragraph on opinion words, adjectives are of particular importance since they are “important indicators of subjectivities and opinions” (Liu and Zhang, 2012, p. 423).

Tagging denotes the “automatic assignment of descriptors, or tags, or input tokens” (Voutilainen, 2003, p. 220). POS taggers assign a tag to each word, and sentiment analysis systems can utilize this information. The tagset by the *Penn Treebank Project*[6] can be considered as standard for POS tagging (Liu, 2012).

Negation and Valence Shifting. Handling *negation* is another crucial issue in sentiment analysis systems. If two sentences are equal except for a single *not*, they potentially express the opposite polarity. A straightforward approach is to encode negation already in the initial features by attaching “NOT” to opinion words which are close to negations in the text. For example, the token *appreciate* in the sentence “I don’t appreciate this behavior” would be encoded with *appreciate-NOT*. Another possibility is to use negations as a second-order feature, i.e. ignore them initially and change the feature representation in a second pass (cf. Pang and Lee, 2008, p. 22).

Unfortunately, negation words like *not*, *no*, or *don’t* do not necessarily indicate a negation. Sentences like the following one require a deeper syntactic analysis: “With the approval of the IRB, BAWAG P.S.K. can use its modern risk measurement methods not only for internal management purposes but also as the basis for its regulatory capital requirements” (BAWAG P.S.K., 2014, p. 160).

Other terms which do not modify the polarity, but the strength of a sentiment word, are *intensifiers* and *diminishers*. Words like *slightly* or *very* weaken respectively strengthen the sentiment. Polanyi and Zaenen (2006) suggest to account for such *contextual valence shifters*[7] by assigning values to sentiment words based on their predecessors. The standard value is +2. If the sentiment word is preceded by a negation, -2 is assigned. Intensifiers are taken into account by assigning +3, diminishers by assigning +1. The authors admit that this method “falls far short of an adequate solution to this problem” (Polanyi and Zaenen, 2006, p. 2), but it is relatively easy to implement and therefore popular.

Vector Space Model (VSM). The vector space model was originally introduced by Salton et al. (1975). It is a popular model for the numeric representation of textual data. The *document space* comprises all documents in the corpus. A *document vector* represents the features f_i of a document. If the document corpus contains m distinct features, a document vector $\overrightarrow{d_j}$ is given by

$$\overrightarrow{d_j} = (f_1, f_2, \ldots, f_m). \tag{3.1}$$

The *feature space* comprises the m distinct features in the document corpus. There are different ways for representing the features: if only feature presence is relevant, it is sufficient to use a bit vector which encodes the absence of a feature by 0, and its presence by 1. Alternatively,

[6]See `http://www.cis.upenn.edu/~treebank/home.html`, accessed June 18th, 2014.

[7]This class of words is also called *modifiers*.

the document vector can be built up by weighted features, for example TF-IDF term weights as introduced in Section 3.3.2. In any case, document vectors are typically very sparse because texts are built up by a relatively small amount of words chosen from a comprehensive vocabulary. For more details about the VSM, see Salton et al. (1975) and Manning and Schütze (1999, p. 539f).

Using the concepts described so far, the document corpus can be represented by means of a feature-document matrix which consists of the individual document vectors. The dimension of this matrix is $m \times n$ if the feature space consists of m elements and the document space of n elements.

3.3.2 Term Weighting

It is safe to assume that not every word in a document is equally important. Hence, merely counting them is not the best indicator for their importance. A better approach is to weigh terms according to a *weighting scheme*. The question is: which term weighting schemes enhance the effectiveness of a specific text mining task? An often cited paper by Salton and Buckley (1988) discusses three essential components of such weighting schemes:

- *Term Frequency (TF):* frequently occurring terms in documents can indicate relevant items. Such *local weights* have been used since decades in content analysis systems. In its simplest form, TF just counts the occurrences of a term in a document.
- *Inverse Document Frequency (IDF):* the downside of term frequency is its poor performance when high frequency terms are omnipresent in the whole collection. Measuring inverse document frequency helps to assess a term's importance within the entire document corpus. It does so by favoring terms concentrated in a few documents. IDF is the most popular *global term weight*.
- *Normalization:* longer documents usually contain more distinct opinion words than shorter ones, but this should not automatically increase their relevance. Hence, an appropriate normalization factor should be used in the text analysis system.

These steps are usually denoted as *TF-IDF*, a technique originally developed in the field of information retrieval. While Kearney and Liu (2014) observe that *TF-IDF* term weighting schemes are very popular in the field of text mining, other researchers express doubts whether this is the best approach for considering differences in relative term importance. Pang and Lee argue that "overall sentiment may not usually be highlighted through repeated use of the same terms" (Pang and Lee, 2008, p. 21). Instead, researchers should look at term *presence* by using a binary-valued feature vector indicating whether a term occurs or not. Nevertheless, this thesis makes use of a proven term weighting scheme which is explained in the next paragraph.

Calculation of Term Weights. Chisholm and Kolda (1999) present a general term weighting formula. It covers three main elements of term weighting schemes, namely local weights, global weights, and document normalization:

$$w_{i,j} = L_{i,j} G_i N_j \tag{3.2}$$

In the equation, $w_{i,j}$ is the weight of term i in document j. $L_{i,j}$ denotes the *local weight* of term i in document j. Its value is always based on the present document and ignores other documents in the corpus. G_i is the term's *global weight*. It considers the term's frequencies in the whole document corpus. Finally, N_j is a normalization factor which takes the document's length into consideration. This is particularly important if the length of the analyzed documents varies.

An established formula for the local weight is given by

$$L_{i,j} = \begin{cases} 1 + \log(tf_{i,j}) & \text{if } tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases} . \tag{3.3}$$

It is presented in Chisholm and Kolda (1999), but also in textbooks like *Foundations of Statistical Natural Language Processing* (cf. Manning and Schütze, 1999, p. 543). The term frequency is denoted as $tf_{i,j}$. The rationale behind the log transformation is that a term which appears 10 times in a document is most likely not 10 times more important than another term which appears only once. Consequently, the impact of high frequency terms is reduced by applying the logarithm function.

The most popular global weight is the *inverse document frequency* (IDF) which has already been explained in the items above. In

$$G_i = \log\left(\frac{N}{df_i}\right), \tag{3.4}$$

the total number of documents is denoted by N, and df_i the number of documents where term i occurs at least once. The IDF can also be described as "the logarithm of the inverse of the probability that term i appears in a random document" (Chisholm and Kolda, 1999, p. 5).

A common approach for *document normalization* is given by

$$N_j = \frac{1}{\sqrt{\sum_{i=0}^{m}(G_i L_{i,j})^2}}. \tag{3.5}$$

It uses the notation by Chisholm and Kolda (1999), but Salton and Buckley (1988) discuss the approach in more detail. The rationale is to create weighted document vectors with equal lengths in order to make them better comparable. This operation is also referred to as *cosine normalization* and results in a vector with a length of one. The term weights used in this equation are the product of the local and the global weight as discussed before.

Another way to conduct length normalization is to divide the local weight by the logarithm of the average count of terms in the document, see Chisholm and Kolda (1999). This paper also gives a general overview on term weighting strategies.

3.3.3 Feature Selection

Merely using all individual words in a document as input for a sentiment analysis system could lead to a poor performance because the employed algorithms would consider too many irrelevant features. Reducing the dimensionality of the feature space is an option to resolve this issue. There are two general approaches: the first one is to filter all irrelevant features, e.g. by utilizing a stop word list. The second one is to find and select specific features, for example the best suited ones for a classification task.

The following paragraphs explain some common techniques for *feature selection* in the field of text mining, focusing on the ones that work well in classification tasks. The information is based on Aggarwal and Zhai (2012a, p. 167ff) and Gupta (2006, p. 112ff).

Stop Words and Lexicon-based Filtering. These are simple, but effective and therefore often used techniques for filtering irrelevant features. *Stop words* are commonly used words which do not carry relevant information for the analysis. Examples are conjunctions, articles or other function words with little meaning. They are usually filtered based on a predefined stop word list. Another method for filtering is to simply remove all features which do not belong to a sentiment lexicon. The remaining features are used in lexicon-based approaches as explained in Section 3.3.4.

Stemming. This is a method for mapping different grammatical forms of a word to their common stem, i.e. multiple features are reduced to a single one. The different forms of the same word appear because of singular and plural forms, different tenses, and other prefixes or suffixes. For example, the word forms *take*, *taken*, and *taking* could be reduced to their common stem *tak*.

Document Frequency-based Feature Selection. While IDF reduces the impact of words which are omnipresent in the whole corpus (see Section 3.3.2), it might be also necessary to filter very *infrequent* terms. If a word only occurs in less than 10 %[8] of the documents in the analyzed corpus, it will most likely be of little use for prediction purposes. Hence, document frequency-based feature selection filters all terms below a specific threshold (cf. Aggarwal and Zhai, 2012b, p. 81f).

Feature Selection With the Use of Class Labels. In classification tasks, it makes sense to utilize class labels for feature selection. The objective is to identify those features which have the highest discriminatory power in a classification problem. In this paragraph, two popular methods for feature selection based on the use of class labels are introduced.

A simple approach is to calculate the *Gini Index*, a measure which is also used for quantifying the distribution of wealth in a country. It is calculated as the sum of squared class probabilities:

$$G(w) = \sum_{i=1}^{k} p_i(w)^2 \tag{3.6}$$

where $G(w)$ is the Gini index for a word w. The term $p_i(w)$ denotes the "conditional probability that a document belongs to class i, given the fact that it contains the word w" (Aggarwal and Zhai, 2012a, p. 168). If one builds a sum of $p_i(w)$ over all k classes where w occurs, the result

[8] 10 % should be seen as an example. The best threshold might differ from this percentage.

is one. This is also the maximum value for $G(w)$. For example, if all documents containing the word *bankruptcy* (b) belong to the class *negative* (n), it holds that $G(b) = p_n(b) = 1$. The minimum value for the Gini index is $1/k$. It occurs when the classes containing w are evenly distributed.

Another method is based on the concept of *information theory* and works with the *entropy* of a feature. Entropy is a measure of disorder or impurity in a data set. A perfectly homogeneous data set has an entropy of zero. In contrast, a completely disordered data set has an entropy of one. If there are k classes, the entropy H of a data set can be computed by

$$H = -\sum_{i=1}^{k} P_i \log(P_i), \tag{3.7}$$

where P_i is the probability that a document belongs to class i. If a feature is able to reduce the entropy in a data set by a large amount, its *information gain* is high. Such features have a relatively high ability to predict the corresponding class.

For calculating the information gain, one has to compute the entropies given the presence or absence of a feature in a data set and subtract the results from the entropy of the original data set.

$$\begin{aligned} I(w) = -\sum_{i=1}^{k} P_i \log(P_i) + F(w) \sum_{i=1}^{k} p_i(w) \log(p_i(w)) + \\ +(1 - F(w)) \sum_{i=1}^{k} (1 - p_i(w)) \log(1 - p_i(w)) \end{aligned} \tag{3.8}$$

where $I(w)$ is the information gain for a given feature respectively word w. The conditional probability $p_i(w)$ is determined in same way as for the Gini index. $F(w)$ is defined as the fraction of documents containing w (Aggarwal and Zhai, 2012a, p. 169).

Both the Gini index and the information gain are popular methods for identifying the most relevant features for classification. They primarily serve as split criteria for *decision trees*, but it is also possible to use them for feature selection.

3.3.4 Lexicon-based Approaches

These methodologies belong to the class of *unsupervised* approaches since they work without class labels and employ words with a known polarity for sentiment extraction. A *sentiment lexicon* is incorporated for determining the degree of positivity or other features of the text (cf. Pang and Lee, 2008, p. 27). Obviously, the lexicon's quality is crucial for this method.

Lexicon-based approaches start with extracting all *features* (see Section 3.3.1) from the given text data. The next step is to annotate the extracted features with their *semantic orientation* as indicated in the underlying dictionary. The latter term is defined as the "polarity and strength of words, phrases and texts" (Taboada et al., 2011, p. 268). The annotated terms are then further processed by techniques like term weighting or negation handling. Finally, the derived sentiment scores are aggregated which results in a score representing the text's overall sentiment (Taboada et al., 2011).

A basic, but very popular approach is to treat documents as a *bag of words* (BoW) when employing lexicon-based approaches. This methodology ignores the document's internal structure, e.g. its word order and grammar. According to Loughran and McDonald (2011), most textual analysis studies in finance utilize the BoW model, often in combination with term weighting schemes.

Creating a Sentiment Lexicon. A sentiment lexicon is built up by labeling text elements according to their semantic orientation. Liu and Zhang (2012) enumerate three possibilities for creating a sentiment lexicon: manual approach, dictionary-based approach, and corpus-based approach. The first one is cumbersome, so it is typically not used separately, but in combination with the other approaches.

The *dictionary-based approach* starts with a small set of *seed opinion words*, which are manually selected and have a known semantic orientation—at least within the context of a specific domain. These words are linked to a dictionary like *WordNet*[9] in order to find synonyms and antonyms. After adding the latter to the lexicon, the next iteration starts. This algorithm is repeated until no more words can be added. Unfortunately, this straight-forward approach has a disadvantage: since general dictionaries are used for expanding the lexicon, domain-specific orientations of opinion words can hardly be found (cf. Liu and Zhang, 2012, p. 429).

A possibility to overcome this issue is the *corpus-based approach*. It is based on a work by Hatzivassiloglou and McKeown (1997) and starts with seed opinion words as well. In contrast to the dictionary-based approach, new words and their orientations are identified via linguistic constraints. For example, if two adjectives are joined via *and*, they usually have the same semantic orientation. The reason for this is that it would be unnatural to write sentences like "The outlook for this bank is uncertain and the financials are solid". In this sentence, one would rather use *but* instead of *and*. Similar *sentiment consistency* rules can be constructed for several connectives. Liu and Zhang (2012) names *or*, *but*, *either-or* and *neither-nor*. However, this technique must be seen as a heuristic since it is possible to break such rules. More fine-grained analyses also distinguish between intra-sentential (within a sentence) and inter-sentential (between neighbors) sentiment consistency.

Popular Lexicons. As already mentioned above, the *General Inquirer* (GI) was among the first popular dictionaries for content analysis. The GI's word lists are closely related to the *Harvard IV-4* dictionaries, which are still used in research. Other popular sentiment lexicons are *SentiWordNet* or *Bing Liu's Opinion Lexicon*[10].

Domain-specific Lexicons. Besides these general word lists, domain-specific dictionaries have been developed and became popular during the last years. Especially in the financial domain,

[9] See `http://wordnet.princeton.edu/`, accessed June 12th, 2014.

[10] For further lexicons and more detailed descriptions, the interested reader is referred to the following website by Christopher Potts: `http://sentiment.christopherpotts.net/lexicons.html`, accessed November 19th, 2014.

this seems to be essential: according to Loughran and McDonald (2011), almost three quarters (73.8 %) of typically negative words[11] cannot be considered as negative when they appear in financial texts. Kearney and Liu (2014) give the examples *tax* and *liability*. These words appear in the *Harvard IV Negative Word List* (H4N), but are neutral when used in a financial context, e.g. in an annual report. Hence, finance-specific dictionaries can improve sentiment analysis in this domain. In 2011, Loughran & McDonald developed such a dictionary (L&M lists). Their list of negative terms in a financial context is with 2,337 words the largest one. They further created lists containing words related to *uncertainty* and *positivity*, which are also relevant for this thesis (cf. Loughran and McDonald, 2011). The L&M lists became quickly popular in the area of sentiment analysis in finance.

For the creation of their lists, Loughran and McDonald analyzed all 10-K annual reports[12] published between 1994 and 2008. Because of data requirements, they had to filter the set of annual reports according to some criteria like length and availability of quantitative data used for regression models (returns, size, book-to-market, institutional ownership). In the next step, the set of documents was parsed into vectors of words and word counts. The authors also removed tables and exhibits and apply the *TF-IDF* term weighting scheme. For the actual creation of their word list, Loughran and McDonald "carefully examine all words occurring in at least 5 % of the documents, to consider their most likely usage in financial documents (including inflections)" (Loughran and McDonald, 2011, p. 12). In order to test for reliability of the word list, they linked them to quantitative variables like returns, trading volume, volatility, or unexpected earnings.

3.3.5 Supervised Classification Approaches

These machine learning approaches use statistical inference for classification. The latter can be generally defined as "the separation or ordering of objects (or things) into classes" (Gupta, 2006, p. 106). Supervised learning approaches require a *training set*. In addition, a *test set* is needed if the constructed model should be evaluated. Both sets are usually taken from the same document corpus. Depending on the type of documents, the data set either has to be labeled manually by researchers respectively professional annotators, or one can use existing class labels. An example for this are online reviews, where authors usually assign a rating of 1-5 stars.

In a next step, an algorithm is trained on the corpus described above. It learns the classification function based on the training set. The trained model is then applied on the test set for determining its performance. Popular supervised classification algorithms in the domain of financial text mining are *Naïve Bayes* (NB) and *Support Vector Machines* (SVMs) (Kearney and Liu, 2014; Takala et al., 2014).

Naïve Bayes. This algorithm belongs to the class of *generative classifiers*. It is simple and one of the most popular approaches for supervised learning. Naïve Bayes uses a probabilistic model and assumes that the different terms are *independently distributed*. This is why it is called *naïve*—it assumes that the occurrence of a word in a document is independent of the presence (or absence) of all the other words in the document respectively the feature vector. This is certainly

[11] I.e. words which are considered as negative when used in everyday language.

[12] 10-Ks are annual reports required by the US Securities and Exchange Commission (SEC).

not the case in a real-world setting. Li (2010) gives an example from financial statements, where the words *adverse effect* and *material*[13] are often used within the same sentence. Fortunately, empirical evidence shows that the assumption of independence hardly affects the results (Li, 2010; Pang et al., 2002). It is further worth mentioning that Naïve Bayes works with the *bag of words* representation of the text corpus, i.e. it ignores the position of words in a document.

The algorithm starts with computing the *posterior probability* for each class. In a next step, the document can be matched to the class with the highest probability (cf. Aggarwal and Zhai, 2012a, p. 181f). Written in formal terms, the approach assigns a document d to the class $c^* = \text{argmax}_c P(c \mid d)$, i.e. the class where $P(c \mid d)$ given argument c reaches its maximum value.

The following explanation of the Naïve Bayes approach is based on Li (2010) and Pang et al. (2002). The basic idea relies on *Bayes' rule*, which was developed by the English statistician *Thomas Bayes* in the 18th century. In general, the rule allows for calculating the probability of event A given event B. In this thesis, it calculates the probability of class c given a document d:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \tag{3.9}$$

Under the assumption of independence among the features, the term $P(d \mid c)$ can be decomposed using the document's set of features, denoted as $\{f_1, \ldots, f_m\}$, as follows:

$$P_{NB}(c \mid d) = \frac{P(c)P(f_1 \mid c)P(f_2 \mid c)\ldots P(f_m \mid c)}{P(d)} = \frac{P(c)\prod_{i=1}^{m} P(f_i \mid c)}{P(d)} \tag{3.10}$$

In the present case, $P(d)$ is constant and does not influence the selection of c^*[14]. Hence, the formula for selecting the document class using Naïve Bayes can be given by:

$$c^* = \text{argmax}_c\, P(c) \prod_{i=1}^{m} P(f_i \mid c) \tag{3.11}$$

While the equations above explain the general approach, specific implementations make different assumptions about the distribution of the features f_i. According to Aggarwal and Zhai (2012a), the following two models are commonly employed:

- *Multinomial Model:* This approach takes *word frequencies* within the documents into account. Like all Naïve Bayes implementations, the multinomial model makes use of the BoW assumption. Hence, the document is split up into a set of terms $Q = \{t_{i_1} \ldots t_{i_m}\}$ and their corresponding frequencies $F = \{f_{i_1} \ldots f_{i_m}\}$. This allows for modeling documents as samples of a multinomial feature distribution. The goal is "to model the posterior probability that the document [...] belongs to class i, given that it contains the terms Q" (Aggarwal and Zhai, 2012a, p. 183). In general, multinomial models assume that the document length is independent of the class. Several variations of this approach exist and include, for example, category hierarchies.

[13] In this example, *material* refers to the accounting concept of *materiality*, see e.g. `http://www.businessdictionary.com/definition/materiality.html`, accessed June 24th, 2014.

[14] Ignoring $P(d)$ makes it impossible to calculate the actual probability value. However, this is not an issue since we are only interested in identifying the class c.

- *Multivariate Bernoulli Model:* This model ignores the frequency of features in a text. It rather recognizes their *presence* or *absence* in a text via a binary feature value. McCallum and Nigam (1998) find by an empirical comparison that the multivariate Bernoulli model achieves good results if the vocabulary size is small, but performs poorly with long texts respectively large vocabularies. For classification tasks with long documents, it is better to use the multinomial model.

Support Vector Machines (SVM). The family of *linear classifiers* aims to construct a linear predictor of the form $p = \overline{AX} + b$. The elements of this equation can be explained as follows (Aggarwal and Zhai, 2012a, p. 193):

- $\overline{X} = (x_1 \dots x_n)$ represents the normalized document word frequency vector
- $\overline{A} = (a_1 \dots a_n)$ is a vector of linear coefficients
- b is a scalar.

If class labels are categorical, the predictor p can be seen as a *hyperplane* used for separating classes. Examples for linear classifiers are *Regression models*, simple *Neural Networks*, and *Support Vector Machines* (SVM).

The latter is described in more detail here. Support Vector Machines are *large-margin classifiers* (Pang et al., 2002) and work with the *Structural Risk Minimization*[15] principle (Joachims, 1998). SVMs aim to find "good" linear separators. Figure 3.1 shows three separating hyperplanes between the classes x and o. In this case, hyperplane A has the largest *normal distance* from the data points. In other words, it provides the maximum *margin of separation* and is therefore better than B and C (cf. Aggarwal and Zhai, 2012a, p. 194).

If the data are not linearly separable in their original representation, it is still possible to employ SVMs by projecting the data into a higher dimension. This is done by applying the so-called *kernel trick*. Explaining this technique in detail would be out of the scope of this thesis, but the interested reader is referred to Alpaydin (2014), who introduces the kernel trick from page 359 on.

Early work on Support Vector Machines shows their particular suitability for text classification tasks. Joachims (1998) concludes from the properties of text data that SVMs are very well suited for this kind of tasks because of the following reasons:

- If machine learning approaches are applied on textual data, the algorithms usually have to deal with large feature spaces. SVMs are known to handle them well because they resist *overfitting*[16]. Other machine learning algorithms (e.g. decision trees) tend to overfit the training data especially when the number of training instances is small compared to the number of features (cf. Gupta, 2006, p. 125).

[15]"The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error" (Joachims, 1998, p. 2).

[16]The term *overfitting* is used when a model deals perfectly with the training set, but fails in classifying the test data. An overfitted model does not describe the underlying relationship, but rather anomalies, noise, or outliers (see e.g. Gupta, 2006, p. 123).

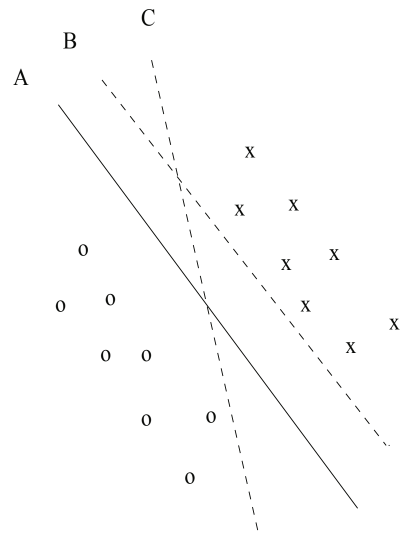

Figure 3.1: Separating hyperplanes between two classes (taken and slightly adapted from Aggarwal and Zhai, 2012a, p. 194).

- SVMs are well suited for the typically sparse feature vectors[17] of documents.
- Text data is usually linearly separable. Hence, SVMs work well in their basic version, i.e. with a linear kernel function.

In addition, Joachims (1998) also provides empirical evidence for their good performance and finds that SVMs generally outperform Naïve Bayes. Support Vector Machines are also common in the financial domain, e.g. for financial distress prediction (Hájek and Olej, 2013)[18].

The simple case depicted in Figure 3.1 separates only two classes. Finding the best separating hyperplane, represented by the vector $\overrightarrow{w}$, can be formulated as a constrained optimization problem.

$$\overrightarrow{w} = \sum_{j} c_j \, \alpha_j \, \overrightarrow{d_j}, \quad \alpha_j \geq 0 \tag{3.12}$$

This defines the hyperplane $\overrightarrow{w}$ based on a training data set consisting of j elements. The variable $c_j \in \{1, -1\}$ represents the correct class[19] of a document vector $\overrightarrow{d_j}$, and α_j is determined by solving a dual optimization problem. For particular document vectors, α_j will be greater than zero. These are called *support vectors* since they solely define $\overrightarrow{w}$. All the other document vectors do not influence the hyperplane. Figure 3.2 illustrates this with a linearly separable case.

[17] The feature vectors of documents are sparse because a single document consist of relatively few different words taken from the whole corpus.

[18] See Section 3.4 for further examples.

[19] In this simple case, the available two classes are denoted as -1 and 1, e.g. *negative* and *positive*.

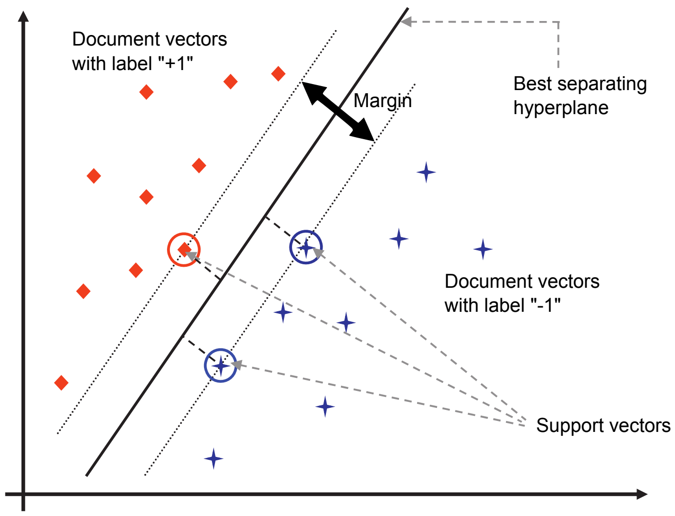

Figure 3.2: Support vectors define the best separating hyperplane (taken and adapted from Kim and Ahn, 2012, p. 1801).

As the figure shows, support vectors are those document vectors which have the closest distance to the best separating hyperplane.

The classification decision for SVM's with two classes is straightforward: one has to decide on which side of the hyperplane the document is placed (cf. Pang et al., 2002, p. 82).

In case of a *multiclass* problem, more than two classes have to be separated. Unfortunately, direct solutions for such problems are complex. Hence, a well-established approach is to tackle them with a combination of binary SVMs.

Kim and Ahn (2012) present such a multiclass SVM technique for *ordinal* classes and use it for predictions of corporate credit-rating. The algorithm is called *Ordinal Multi-Class Support Vector Machines* (OMSVM) and incorporates *Ordinal Pairwise Partitioning* (OPP), an approach for improving ordinal multiclass classification. OMSVM consists of the *preparation* and the *interpretation* phase. The former comprises the division of the training set into $k-1$ groups and constructs a binary SVM classifier for each. The interpretation phase assigns classes to the input data by fusing the binary classifiers.

Figure 3.3 gives an example for OMSVM classification using the method *One-Against-The-Next* during the preparation phase and *Forward fusion* in the interpretation phase[20]. During the preparation phase, the method *One-Against-The-Next* trains $k - 1 = 3$ binary SVM classifiers, in particular $(1vs2)$, $(2vs3)$, and $(3vs4)$. In the interpretation phase, the previously constructed SVM models classify the input data. The *Forward fusion* approach begins with $(1vs2)$ when fusing the classifiers. If this model decides in favor of class 1, the respective document is assigned to class 1. Otherwise, model $(2vs3)$ is applied. In case of a decision for class 2, the document can be assigned to this class since the previous model already excluded class 1. If

[20] Alternatively, one could also use *One-Against-Followers* and *Backward fusion* or a combination of the methods.

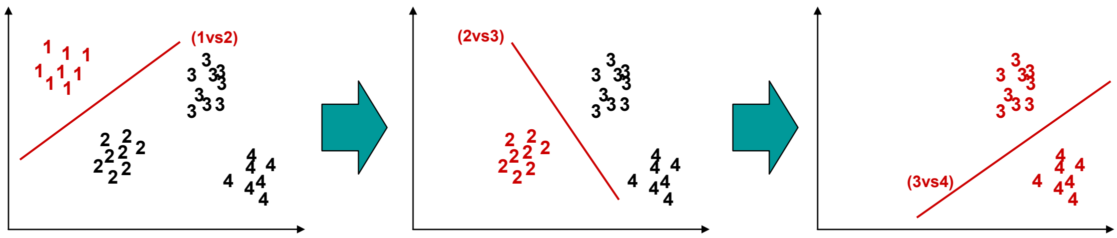

Figure 3.3: OMSVM using the One-Against-The-Next + Forward approach (taken and slightly adapted from Kim and Ahn, 2012, p. 1806).

$(2vs3)$ decides in favor of class 3, the last model $(3vs4)$ is used for determining whether the document belongs to class 3 or 4.

In their conclusion, Kim and Ahn state that “OMSVM is an effective and efficient classifier for solving ordinal multi-class classification problems” (Kim and Ahn, 2012, p. 1810).

3.4 Related Work

Sentiment analysis in general and its application in the financial domain in particular gained a lot of interest within the last decade. Although there seems to be no study about risk sentiment analysis in banking, there is a considerable number of closely related papers. This section discusses relevant studies and focuses on the problems they tackled, their key findings, text data sources, and the applied methodologies.

A good introduction to this field of research is the article “Textual sentiment in finance: A survey of methods and models” (Kearney and Liu, 2014). Even though this survey does not only focus on risk sentiment analysis, the presented approaches and data sources are definitely relevant for this work. Some, but not all of the papers discussed in this section are also mentioned in the survey by Kearney and Liu.

3.4.1 Overview

Related studies deal with risk sentiment analysis in manifold ways. A common question tackled by researchers is whether corporate disclosures drive stock price volatilities or future earnings of the respective firm (Groth and Muntermann, 2011; Kogan et al., 2009; Li, 2006; Tsai and Wang, 2013). Hence, they focus on the risk an *investor* takes if he or she buys stocks of a company. Generally spoken, these studies find significant correlations between sentiment extracted from corporate disclosures and future volatilities. In particular, Groth and Muntermann (2011) find that intraday market movements are driven by corporate disclosures and news stories. All other papers analyze year-long periods since they employ annual reports as text data source: Kogan et al. (2009) use past volatility as a baseline and show that incorporating text information significantly enhances the model’s predictive power. Li (2006) finds that the frequency of words related to *risk* and *uncertainty* negatively correlates with future earnings.

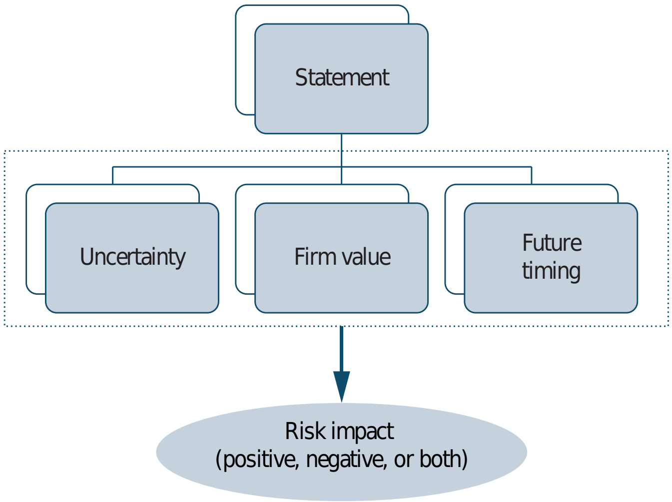

Figure 3.4: Concept for risk-related statement recognition (taken and slightly adapted from Lu et al., 2010, p. 80).

Other papers deal with *financial distress prediction*, for example Hájek and Olej (2013). In this study, the authors treat financial distress as a two-class problem by using *Standard & Poor's* ratings, in particular investment grade (IG) and non-investment grade (NG)[21]. As a baseline, they classify companies based on various financial indicators[22]. It turned out that the inclusion of sentiment indicators improved financial distress prediction. Henselmann and Scherr (2012) have a similar goal. They try to predict bankruptcy by analyzing annual reports of American companies. As a warning sign for bankruptcy risks, the authors use specific tags from XBRL[23] taxonomies. Such tags add semantics to the textual content of semi-structured XBRL reports.

Besides these predictive approaches, a third group of studies in the field of risk mining tries to improve the *identification* of risks. Lu et al. (2010) develop methods for the automatic detection of risk-related statements in textual data. According to their definition, such a statement "imparts information that could affect investors' beliefs about a firm's future value" (Lu et al., 2010, p. 79). It has to be (1) uncertain, (2) related to firm value, and (3) future centric. Figure 3.4 depicts their concept for risk-related statement recognition.

Leidner and Schilder (2010) aim to detect risks before they result in adverse events. They employ *information extraction* (IE) and *web mining* for creating company specific *risk maps*, which aggregate and visualize the extracted risk types. In their conclusion, Leidner and Schilder state that they addressed two general research questions: first, they aimed to overcome a weakness of frequency-based methods, namely to identify rare, but particularly harmful risks. Second, they use web mining to "bridge the vocabulary gap, i.e. how to match up terms and phrases in financial news prose with the more abstract language typically used in talking about risk in

[21] IG bonds have a low default risk, NG bonds are considered as risky.

[22] In particular, they used 13 variables like *return on equity*, *dividend yield*, or *earnings per share*.

[23] XBRL is the XML-based *eXtensible Business Reporting Language*.

general" (Leidner and Schilder, 2010, p. 59). Generally, the work by Leidner and Schilder aims to assist decision makers in the financial sector with the identification and mapping of risks.

3.4.2 Data Sources

The majority of related papers analyzes 10-K annual reports of US companies, available at the US Securities and Exchange Commission's EDGAR system[24], as source for text data (Hájek and Olej, 2013; Kogan et al., 2009; Li, 2006; Loughran and McDonald, 2011). EDGAR provides free access to millions of 10-Ks and is therefore a very well suited text source for risk sentiment analysis. Unfortunately, there is no European counterpart to EDGAR where one could easily download annual reports of banks supervised by the SSM.

Hájek and Olej (2013) use 10-Ks as well as quantitative data from 520 US companies. These input figures date from 2010, whereas the *Standard & Poor's* investment grade data are from 2011. Kogan et al. (2009) employ a much larger data set. The authors collected over 54,000 reports published between 1996 and 2006. In order to exclude irrelevant information, they focus on Section 7 of the 10-Ks, which contains important forward-looking content. In particular, this section comprises the *management's discussion and analysis* (MD&A). Kogan et al. extract this section from the annual report with help of a Perl script. Li (2006) retrieved all 10-Ks published between 1994 and 2005, but excluded reports from financial firms[25] and those without risk-related statements. Loughran and McDonald (2011) utilize 10-Ks from a 14-year period for testing the reliability of their word lists.

Groth and Muntermann employ different data sources, in particular news stories and stock price data. They focus on regulatory-induced corporate disclosures since this type of news seems to cause abnormal stock price reactions. In particular, the authors used a set of disclosures released between August 2003 and July 2005 in order to fulfill article 15 of the German Securities Trading Act. Furthermore, stock price series provided by *Thomson Reuters DataScope Tick History* were included in their analysis.

The analyses of Leidner and Schilder (2010) are based on 170,000 earning calls transcripts from the *StreetEvents* database. Earning calls are regular events where managers report about the company's current situation and answer questions from business analysts.

Finally, Lu et al. (2010) collected firm-specific news articles from the *Wall Street Journal* for their experiments. In total, their study incorporates 1,529 sentences.

For a more comprehensive discussion of potential data sources, the reader is referred to Kearney and Liu (2014). In Chapter 2 of their survey, they name three main sources of data in the field of textual sentiment in finance: corporate disclosures, media articles, and internet messages.

3.4.3 Methodologies and Processing Pipelines

Analyzing the methodologies in related studies shows that the authors work with similar approaches for extracting sentiment from texts. Linguistic preprocessing generally involves to-

[24] See `http://www.sec.gov/edgar/searchedgar/webusers.htm`, accessed December 3rd, 2014.

[25] According to Li, risk-related word have different implications for financial firms.

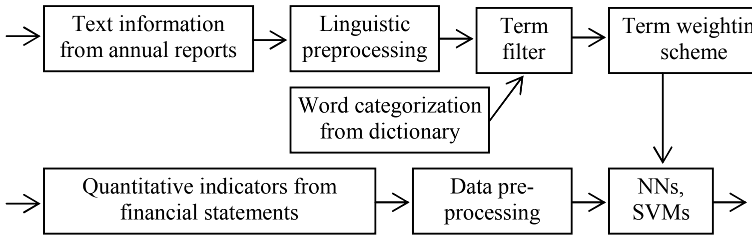

Figure 3.5: Research methodology for financial distress prediction (taken from Hájek and Olej, 2013, p. 4).

kenization, lemmatization, and removing nonessential items like tables, exhibits, or digit sequences. In almost every study, the authors also make use of term weighting schemes. With the selected features and additional quantitative data, the authors either train machine learning algorithms, or use the data for regression analyses. The following paragraphs explain the individual approaches in a more detailed way.

Groth and Muntermann (2011) start with labeling the documents using an intraday market risk model called *realized volatility*. Feature extraction is done with a *String Tokenizer* as well as a *Porter Stemmer*. Several filters reduce the dimensionality of the feature set, and a *TF-IDF* weighting scheme modifies the representation of the features. For classifying the documents, the study employs different supervised learning approaches: Naïve Bayes, k-Nearest Neighbor, Neural Network, and Support Vector Machines. Finally, the authors evaluate their model using the measures *accuracy*, *recall*, *precision*, and the *F-measure* which combines precision and recall. In addition to this classic model evaluation, a simulation-based evaluation is conducted.

Kogan et al. (2009) work with basically the same preprocessing and feature representation as Groth and Muntermann (2011). In addition to the classic *TF-IDF*, they also apply *LOG1P*, which dampens word frequencies using a logarithmic function. These features were employed to train different regression models by utilizing *Support Vector Regression* (SVR). Stock returns usually change gradually, so past volatility is a very good predictor for volatility in the next period. Hence, the authors use past volatility as a baseline and show that incorporating text information significantly enhances the predictive power.

The research methodology used by Hájek and Olej is depicted in Figure 3.5. After collecting the data, tokenization and lemmatization is done in a preprocessing step. The resulting features are categorized by incorporating the financial dictionaries developed by Loughran and McDonald (2011). In a next step, it is thus possible to construct sentiment indicators by counting and weighting negative, positive, uncertain, litigious, strong modal, and weak modal terms. Finally, both sentiment and financial data are analyzed by a *Neural Network* as well as by a Support Vector Machine.

Lu et al. (2010) also employ supervised learning approaches for their study. In particular, Lu et al. compare *Support Vector Machines* (SVMs) and *Elastic-net Logistic Regression* (ENET) and find that these techniques achieve similar performance levels. POS tags are used for detecting risk-related language, which is identified by syntactic aspects related to uncertainty and future-orientation. For assessing the performance of their approach, Lu et al. calculate the classical measures *recall*, *precision*, and the *F-measure*.

Li (2006) measures risk sentiment in a very simple way: after some basic preprocessing steps, the author just counts the frequencies of the words *risk* and *uncertainty*[26]. He argues that a more sophisticated solution would require better knowledge about context and linguistic structure, and acquiring this knowledge would be too costly. For measuring correlation between risk sentiment and future earnings, Li (2006) conducts regression analyses and several statistical tests for robustness of the results.

Finally, Leidner and Schilder (2010) follow a very individual approach for detecting risk-related statements. The first part of their system starts with abstract descriptions of risk, which they use as basis for retrieving additional candidates from the web. After this step, they relate these candidates "to risk types via a transitive chain of binary IS-A relations" (Leidner and Schilder, 2010, p. 56). The system's second part is a *Risk Monitor*. It utilizes the taxonomy built up in step 1 for analyzing text data, in their case earning calls transcripts. After visualizing the data with a *Risk Mapper*, analysts can employ the tool for identifying concrete risks (cf. Leidner and Schilder, 2010, p. 56f).

3.4.4 Conclusions

The studies presented in this section provide a good basis for this work. Although none of the papers focused on risk sentiment analysis in the banking industry, major parts of their processing pipelines and approaches can be reused. This starts with selecting appropriate data sources. Analyzing annual reports is very popular in this field of research, making this kind of data source an interesting option for the experiments of this thesis. In this context, the approach used by Kogan et al. (2009) is of particular interest. In contrast to other researchers, the authors do not use the whole annual report, but extract just the section with the most important forward-looking text data in it.

Linguistic preprocessing and feature representation follow a similar pattern in the related papers, and the methodologies are also suitable for this thesis. In contrast, it is interesting to see how many different ways are used for feature selection. The chosen approach mainly depends on the author's understanding of risk. The range spans from the simple approach of counting the root words *risk* and *uncertainty* to sophisticated methods of risk detection via taxonomies. A reasonable method is applied by Hájek and Olej (2013), who incorporate the topic-specific word lists by Loughran and McDonald (2011) as term filter. They find that financially distressed firms use a "more negative, less positive, more uncertain, less litigious and more modal tone" (Hájek and Olej, 2013, p. 6). Focusing merely on uncertainty, like Lu et al. (2010) do, would probably be too restrictive for the purpose of this thesis.

Regarding the machine learning algorithms and the incorporation of quantitative indicators, this thesis will work with approaches similar to Groth and Muntermann (2011), Kogan et al. (2009), and Hájek and Olej (2013). All of them define the document labels based on suitable quantitative indicators. For labeling, the studies have to consider the fact that text data are forward-looking, but quantitative indicators reflect the past. Hence, the indicators are taken

[26]The author also includes variations like *risky*, *risks*, *uncertain*, and *uncertainties*.

from one period after publication of the text data. The labeled data is then used for training machine learning algorithms. Particularly interesting is the approach by Kogan et al. (2009), who incorporate quantitative indicators from the previous period as baseline, and Hájek and Olej (2013) with their approach to select features based on a domain-specific lexicon.

Loughran and McDonald (2011) note an important aspect in their conclusion which should also be mentioned here: textual analysis is not able to unveil *causal links* between the tone of publications and quantitative figures like returns. Nevertheless, it helps to explain the impact of information published in the analyzed documents on financial indicators.

CHAPTER 4

Data Sources and the Concept of Risk

After discussing the economic and technical background in a broad manner, this chapter prepares the actual experiments and evaluations by choosing appropriate data sources which reveal information on the banks' risk sentiment. It further investigates the concepts of risk in banking. This is done by presenting the most important quantitative risk measures as well as findings from research on risk culture in banks.

4.1 The Concept of Risk in the Financial Industry

In general, *risks* can be defined as "uncertainties resulting in adverse variations of profitability or in losses" (Bessis, 2002, p. 11). This definition already anticipates the subtle difference between risk and *uncertainty*. While the latter is about the randomness of outcomes, risk denotes the adverse effects of these outcomes.

Every financial transaction generates some degree of *individual risk*. If these transactions are bundled in a portfolio, the risks diversify. For statistical reasons, the portfolio's total risk is always less or equal than the sum of the individual risks. In particular, it depends on the correlations between the individual transactions[1]. Although diversification reduces the risk of a portfolio, losses can still be high if the portfolio is *concentrated*. This is the case when individual transactions within a portfolio are too large and hence strongly influence the portfolio's risk.

The banking industry distinguishes several types of risks (see Figure 4.1). This thesis addresses risk in the banking sector in a generalized way. Nevertheless, the most important types should be briefly discussed (cf. Bessis, 2002, p. 13ff):

- *Credit Risk:* a main business of banks is to issue credits. Unfortunately, there is always the risk that counterparties are not able to fulfill their payment obligations. Reasons for this

[1] If all individual transactions are perfectly correlated, the portfolio's risk equals the sum of the individual risks. Theoretically, the individual risks could also offset each other resulting in a portfolio risk of zero.

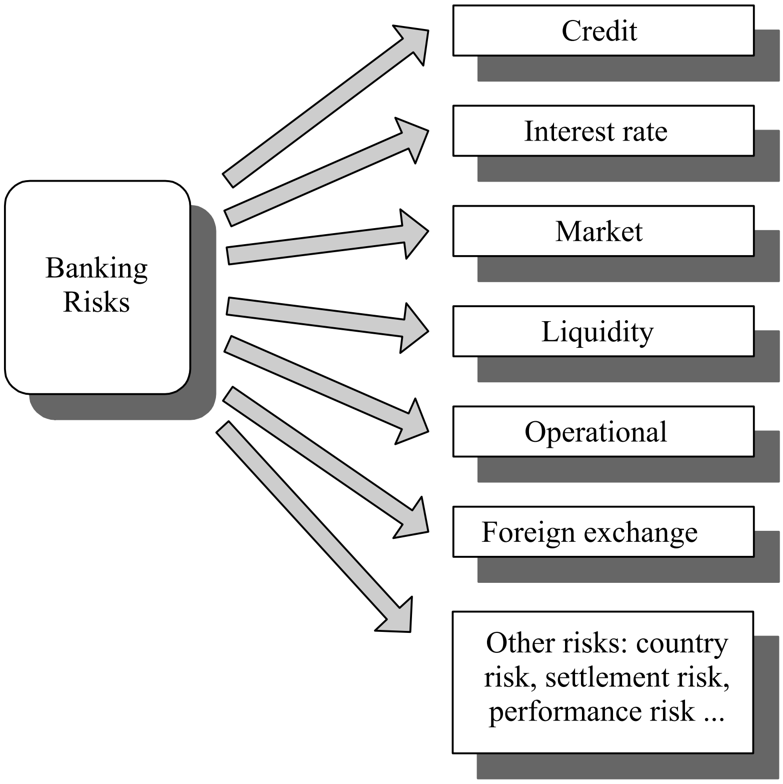

Figure 4.1: Main banking risks (taken from Bessis, 2002, p. 12).

are defaults of individual borrowers[2], but also concentration risks in portfolios or country risks[3]. Credit risk is considered as the most important risk in banking.

- *Interest Rate Risk:* banks lend and borrow money. This leads to revenues and costs depending on variable interest rates. There are two types of interest rate risks: on the one hand, lenders could lose money when interest rates decline. On the other hand, borrowers face higher costs in case of rising interest rates.

- *Market Risk:* this kind of risk refers to adverse deviations of the instruments' market prices in a bank portfolio. An example for market risk is the threat of falling stock prices. The severity of market risk depends on the liquidation period, i.e. the minimum time needed to sell an instrument. The longer this period is, the higher are potential losses due to market movements.

[2]In this context, default stands for the inability of a borrower to pay his or her debt.

[3]Country risk is also called sovereign risk and comprises risks originating from central banks, government sponsored banks, and the legal environment.

- *Liquidity Risk:* if an institution is not able to gather funds at all or if funding is too expensive, it faces illiquidity. In extreme cases, illiquidity leads to failure. The liquidity or funding risk is particularly high when the markets mistrust the institution.

- *Operational Risk:* these risks result from problems within a bank. Operational risks comprise malfunctions in internal systems like IT or reporting, failed internal processes, mistakes by management and employees, and the like.

- *Foreign Exchange Risk:* usually, banks also hold assets and liabilities quoted in foreign currencies. Exchange rates between domestic and foreign currencies change over time, which leads to potential losses.

These items show that risk is omnipresent in the financial sector. Generally spoken, the following holds: the more risk a bank takes, the higher are the potential profits, but also the potential losses. Hence, banks will act—within the regulatory framework—according to their *risk culture* discussed in the next section.

4.1.1 Risk Culture

The financial crisis highlighted the importance of a sound risk culture and better risk management in financial institutions. Since then, several frameworks addressing these issues have been developed, see e.g. Financial Stability Board (2014), Power et al. (2013), or Institute of International Finance (2009). Some authors even argue that there has been an "explosion of interest in risk culture in financial organisations since 2008 as being symptomatic of a desire to reconnect risk-taking and related management and governance processes to a new moral narrative of organisational purpose" (Power et al., 2013, p. 4). The term *risk culture* is driven by the individuals' risk sentiment and can be defined as "the norms and traditions of behavior of individuals and of groups within an organization that determine the way in which they identify, understand, discuss, and act on the risks the organization confronts and the risks it takes" (Institute of International Finance, 2009, p. AIII.2).

With regards to this thesis, it is interesting that the studies cited in the previous paragraph view the language respectively the word choice used by managers as critical since it influences the bank's risk culture. Managers have to ensure that the risk culture sustains different market environments and that "risk management is not being overpowered by revenue generators" (Institute of International Finance, 2009, p. AIII.7).

The *Financial Stability Board* (FSB)[4] published a framework for assessing risk culture in April 2014. It aims to help supervisors understanding the norms, attitudes and behaviors of financial institutions concerning risk. The FSB identifies the following indicators of a sound risk culture (Financial Stability Board, 2014, p. 3f):

[4]The FSB (see `http://www.financialstabilityboard.org/`) is an international organization and promotes effective regulatory, supervisory and other policies. Members of the FSB include the World Bank, the ECB, and the European Commission.

- *Tone from the top:* the language used by senior executives influences a bank's risk culture. Employees will partially adopt their attitudes and behavior. Hence, the managers' word choice is an important factor in this context.

- *Accountability:* employees should be accountable for actions which affect risk-related goals of the institution. Prerequisites for this are the establishment of a policy of risk ownership and clear responsibilities.

- *Effective communication and challenge:* facilitating effective communication promotes an open and critical dialogue within the financial institution. This makes it easier to identify and handle risk issues.

- *Incentives:* there should be incentives for employees that encourage sound risk taking. It is important that such incentives promote long-term interests like *sustained* profitability over short-term revenue generation.

4.1.2 Quantitative Risk Indicators

One of this thesis' aims is to test for relations between textual risk sentiment and quantitative risk indicators. There are various approaches for measuring risk in a quantitative way. For example, the European Banking Authority (EBA) evaluated 53 indicators for its most recent risk assessment report (European Banking Authority, 2014b). Most of these indicators are backward-looking and the underlying data are a couple of months old when the indicators are published. Thus, they are complemented with forward-looking methods like surveys.

As Bessis (2002) remarks, quantitative risk measures are not able to capture all uncertainties. The intangible nature of risk requires to make assumptions and to define scenarios. Despite these limitations, quantitative techniques are the most common approach for risk assessment in the financial industry.

The quantitative risk indicators for the comparisons in this thesis have to fulfill the following criteria:

- The selected quantitative risk indicator has to be a suitable measure for the purpose of this thesis, i.e. it has to represent financial health and the general risk exposure of a bank within a specific period or at a specific point in time.

- For practical reasons, the figures of every analyzed bank have to be publicly accessible. Furthermore, they have to be available for every analyzed bank and for the whole analyzed period. Ideally, the figures should be accessible via financial data providers like *Bankscope* or *Bloomberg* in order to ease retrieval.

- For meaningful comparisons over time, they have at least to be published annually, and at least since 2002.

- The figures have to be comparable among the different banks. This can be reached through standardized methods of calculation.

Banking experts from *Deloitte Financial Advisory* and *The Boston Consulting Group* recommended investigating the indicators described in the following paragraphs since they potentially meet the criteria set out above.

Risk-Weighted Assets (RWAs). This figure represents a bank's total exposure to risk by calculating the risk-weighted amount of its assets. As explained in the *Financial Times Lexicon*[5], RWAs are commonly used for calculating capital adequacy ratios. These are needed for regulations like *Basel III*, which requires banks to hold a minimum capital ratio of 7 % of their risk-weighted assets.

The actual calculation is done by allocating the exposures to several *asset classes* with different weights. Low-risk asset classes are multiplied by a low factor, and vice versa. In case of risk-free assets, the factor is zero. An example for the latter are government bonds, but critics argue that such bonds should be risk assessed as well (cf. Korte and Steffen, 2014).

A standardized approach for assigning the risk weights was introduced by the *Basel II* regulation, which demands ratings from an *External Credit Assessment Institution* (ECAI). These serve as basis for determining the risk weights for the different asset classes, for example corporate exposures or mortgages (Cannata et al., 2012).

For the comparison of different banks, the RWA would have to be represented as ratio of RWA to the bank's size.

Tier 1 Capital Ratio (T1). This indicator is one of the most important ratios based on risk-weighted assets. In particular, it refers to Tier 1 capital as a percentage of RWA:

$$\text{Tier 1 Capital Ratio} = \frac{\text{Tier 1 Capital}}{\text{Risk-Weighted Assets}} \tag{4.1}$$

Tier 1 capital is considered as "best form of bank capital"[6] and has to fulfill several criteria making it relatively secure. As Cannata et al. (2012, p. 12) put it, this ratio "measures the ability of the bank to absorb losses". If the Tier 1 capital ratio is high, the bank acts conservatively and with a high risk buffer. A high ratio can be achieved by either increasing the Tier 1 capital or by reducing the amount of risk-weighted assets, i.e. reducing the amount of total assets or replacing them with safer ones.

This ratio also played a major role during the 2014 EU-wide banking stress test, which was an important part of the preparation phase for the Single Supervisory Mechanism. The stress test had the purpose to assess the resilience of 123 banking groups from 22 EU countries in different macroeconomic scenarios, measured by the impact on the T1. Two different scenarios were used: in the *baseline scenario*, the minimum Tier 1 capital ratio (*hurdle rate*) was set to 8.0 %. The hurdle rate of the *adverse scenario* was set to 5.5 % (cf. European Banking Authority, 2014a).

[5]See `http://lexicon.ft.com/Term?term=risk_weighted-assets`, accessed September 20th, 2014.

[6]See `http://lexicon.ft.com/Term?term=tier-one-capital`, accessed September 21st, 2014.

CDS Spreads. A *Credit Default Swap* (CDS) is a popular credit derivative for transferring credit risk to another party. Hence, it can be considered as an insurance against losses from so-called *credit events* like bankruptcy of the borrower. A *CDS spread* refers to the premium paid by the buyer of this insurance. This premium represents the market participants' perception of the bank's risk (Casu and Chiaramonte, 2011).

In a perfect market with a high number of participants and full information, CDS spreads would therefore be a very good risk indicator. In practice, this holds only for the largest banking groups in the world. According to the interviewed experts, one could name *Deutsche Bank* as an example. However, the CDS market for Austrian banks is too illiquid[7] for producing a meaningful indicator.

(Stressed) VaR. *Value-at-Risk* (VaR) measures the "maximum loss at a preset confidence level" (Bessis, 2002, p. 87). Although not without flaws, VaR is one of the most important concepts for assessing market risk in finance. It is used by both authorities and private firms. Simple forms of VaR rely on historical data and assume that returns are normally distributed. For a EUR 1 million portfolio, a typical statement after conducting a VaR analysis could be: "The investor faces a 1 percent chance of losing EUR 23,260 or more on her index fund on any given trading day" (cf. Chen, 2014, p. 187).

The financial crisis showed that traditional VaR models perform poorly during exceptional periods. In 2009, the *Basel Committee on Banking Supervision* (BCBS) came up with an improvement of the model. Basically, the committee suggested to train the VaR models with historical data from periods of significant financial stress. Besides three other new elements, stressed VaR was added to the Basel framework (cf. Bank for International Settlements, 2009, p. 12), but the figures are currently only available for a small set of banks.

Volatility and Sensitivity. While volatility measures "variations around the average of any random parameter or target variable", sensitivity "captures the deviation of a target variable due to a unit movement of a single market parameter" (both quotes from Bessis, 2002, p. 77). Both measures are incorporated in VaR models, where they capture variations of asset values.

These basic measures are mentioned here because related papers use volatility as a measure of risk (see discussion of related work in Section 3.4). This makes sense if one is interested in the stock price movements of a company, e.g. an investor. In this case, high volatility makes the investment risky because of the high potential losses. For banking supervisors, however, the other risk measures mentioned in this section are much more meaningful.

4.2 Sources for Text Data

Having analyzed potential quantitative risk measures, the next step is to choose suitable data for the text corpus. Like the quantitative measures, the text data have to fulfill some basic criteria:

[7] In this context, illiquidity stands for a relatively low number of transactions.

- The documents' contents have to be suitable for the purpose of this thesis, i.e. (1) they should be published by the bank and (2) they have to contain forward-looking and subjective information about the bank's attitude and expectations towards risk.
- For practical reasons, the documents have to be publicly accessible and available for all banks.
- They have to be published at least annually, and it should be possible to assign the information to a specific period.
- The documents have to be available in the English language.

The following part of this work discusses different sources for potentially suitable text data.

4.2.1 Text Data from Periodic Reports

Like other large companies, banks publish comprehensive reports at least annually. These reports contain information for various stakeholders interested in the bank. In particular, they typically contain the latest financial statements, letters from chairmen and the supervisory board, a management report, a risk report, and disclosures due to regulatory requirements. *Annual reports* are considered as the most important periodic reports. They are the most comprehensive ones and get the highest attention. The following paragraphs describe potentially interesting parts from annual reports:

Letters from CEOs and the Supervisory Board. These carefully crafted documents are a very important part of banks' annual reports and contain valuable information about the management's opinions about risk. Amernic et al. (2010) recognize in their study from 2010 that the word choice of managers strongly influences companies. CEO letters are a way for managers to communicate their attitudes and values. In their conclusion to a chapter about software-based assessments of CEO letters, the authors state that analyzing these documents "could disclose valuable information about such issues as attitudes towards risk, before the organisations concerned stumble into avoidable crises" (Amernic et al., 2010, p. 140).

In a recent working paper, Boudt and Thewissen (2014) study sentiment in CEO letters. In addition to the polarity of individual words, they also pay attention to the position of words in a text. After studying the influence of sentiment in CEO letters on future firm performance, the authors conclude that CEO letters are "important documents for scrutiny" (Boudt and Thewissen, 2014, p. 26).

A potential pitfall of CEO letters is that they usually address both the last period as well as the future. Hence, it could be necessary to extract the forward-looking paragraphs within the documents. An example from such a paragraph reads as follows:

> "This year will, however, be another very difficult one—for the global economy, for the financial sector around the world and also for Commerzbank. We have to face the threefold challenge of coping successfully with difficult conditions, driving forward the integration of our two banks, while at the same time focusing still more

> on meeting, in all that we do, the needs of our clients" (Commerzbank AG, 2008, p. 11).

Risk Reports. These usually lengthy chapters of annual reports describe the bank's risk management approaches and frameworks. In particular, they address the different types of risk as explained in Section 4.1.

Risk reports lack subjective, forward-looking information. Such reports are a purely objective description of the bank's risk management approaches, supplemented with related charts and figures. A typical sentence from a risk report reads:

> "Operations in the divisions that are responsible for credit risk are set up to include a functional risk management specialisation for the commercial and institutional customer segment and the retail and SME segment" (BAWAG P.S.K., 2014, p. 159).

Management Report. Another typical section in periodic publications of banks is the *(interim) management report*, also called *review of operations* or *Director's report*. It is basically a textual summary of the bank's results, its business environment, and regulatory as well as internal developments. However, the last part of the management report is the most interesting one for this thesis since it is an outlook on the next year. Banks write about the expected macroeconomic environment, management guidelines and priorities for the next period. This outlook might be less subjective compared to the CEO letter, but it is often more comprehensive and contains interesting forward-looking information about the bank's attitudes and beliefs regarding the next period.

For example, the *Outlook and opportunities report* of the German *Commerzbank* within their Annual Report 2008 spans over seven pages. Given the difficult conditions at the time when this report was published, the *general statement on the outlook for the Group* sounds quite pessimistic:

> "Due to the ongoing severe market turbulence and the extremely volatile general environment in which we operate, it is currently impossible for us to make any well-founded forecasts for the 2009 results" (Commerzbank AG, 2008, p. 123).

The management report in European publications is comparable to the *Management Discussion & Analysis* (MD&A) in American 10-Ks, i.e. annual reports as requested by the *Securities and Exchange Commission*.

4.2.2 Text Data from Other Sources

Obviously, there are also potentially relevant text sources beyond periodic reports. Examples are press releases, publications in the course of investor relations, and text data from social media.

Internal CEO memos. Since managers have a considerable influence on a bank's risk culture, it would be interesting to use internal memos[8] from CEOs or other top executives as text source

[8] Such memos could be E-Mails, letters, or other kinds of messages.

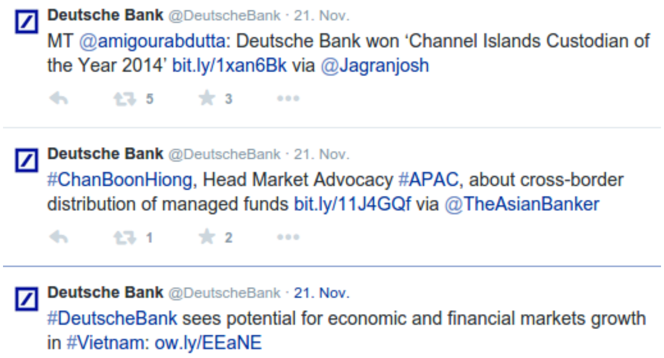

Figure 4.2: Tweets by *Deutsche Bank* (Screenshot from *twitter.com*, taken on November 25th, 2014.

for sentiment analysis. Unfortunately, such memos are not publicly accessible and potentially not written in English.

Press Releases. Banks regularly publish news addressed to the public or to investors on their websites. Press releases focus on communicating public relations, for example new products, social initiatives, sponsoring, and changes in the management board. Hence, the contents of these documents do not seem to be well suited for risk sentiment analysis.

Investor Relations. News published in the course of *investor relations* (IR) are more interesting. The publications do not appear at specific intervals, but relatively often. Regarding the contents, IR news focus on similar topics as the periodic reports do, but in a more condensed manner. For example, when *Erste Bank* published their financial results for the first quarter of 2014, they commented on the figures in an IR news piece[9]. These comments also included justifications for the figures and the expected future trend, but in total, there is only little information regarding future risks.

Social Media. Most banks also have a presence on social media channels. *Facebook* and *Twitter* are the most popular ones, but large banks like *Deutsche Bank* also operate channels on *YouTube*, *Flickr*, *visual.ly*, and *Google+*[10]. Within these platforms, the bank runs various channels for different interest groups and regions.

The rising popularity of social media results in large amounts of text data, which is in turn a valuable source for sentiment analysis. The review article *Techniques and Applications for Sentiment Analysis* states that "Twitter and Facebook are a focal point of many sentiment analysis applications" (Feldman, 2013, p. 87).

[9] See `https://www.erstegroup.com/en/Investors/News/Investor-News/2014/4/30/EN~Investor-News`, accessed November 25th, 2014.

[10] For an overview, see `https://www.db.com/medien/en/content/socialmedia.htm`, accessed November 25th, 2014.

However, a bank's social media channels are of limited usefulness for risk sentiment analysis. Similar to press releases, the contents of such publications comprise a wide range of topics, for example public relations or general news about the financial sector and the world economy. Figure 4.2 illustrates this.

4.3 Conclusions

This chapter deals with risk in the financial industry. After a general definition and an explanation of the different risk types, a section about risk culture in banks argues that the *tone from the top* is an important factor which shapes the risk culture of a financial institution. In this conclusion, both the quantitative risk indicators introduced in Section 4.1.2 and the text data sources introduced in Section 4.2 are discussed with regards to the respective criteria.

4.3.1 Discussion of Quantitative Risk Indicators

Table 4.1 gives an overview whether each quantitative risk indicator satisfies the four criteria described in Section 4.1.2. The following paragraphs provide a more detailed discussion.

		Criteria			
		Suitable measure	Accessible & available	Publication frequency	Comparable
Indicators	CDS Spreads	✗	✗	✓	✗
	Risk-Weighted Assets	✓	✓	✓	✗
	Tier 1 Capital Ratio	✓	✓	✓	✓
	(Stressed) VaR	✗	✗	✓	✓
	Volatility, Sensitivity	✗	✗	✓	✓

Table 4.1: Overview on the assessment of quantitative risk indicators
Note: ✓ indicates *applicable*, ✗ stands for *not or just partially applicable*

Suitability. Both RWAs and the T1 are suitable measures, i.e. they represent the bank's financial health and its general risk exposure. CDS spreads would not be suitable for all banks in the test set, but only for the largest ones. Value-at-risk models primarily quantify *market risk*, so the numbers do not represent the bank's general risk exposure. Finally, volatility and sensitivity are not suited for the purpose of this thesis because they only measure the deviation of target variables like stock prices or asset values.

Accessibility and Availability. CDS spreads are not published in a bank's annual report, so one would have to retrieve the figures from financial data providers, which are usually not publicly accessible. Stressed VaR was only introduced in 2009, so the data are not available for the whole period of time. The availability of volatility and sensitivity depends on the target variable. Both RWAs and the T1 are published within annual reports, so they satisfy this criterion.

Publication Frequency. All of the discussed indicators are either continuous measures or published in sufficiently short time intervals.

Comparability. CDS spreads are in essence market prices, but their comparison is only meaningful for very large banks. Risk-weighted assets are calculated by means of standardized methods, but they do not completely fulfill the criterion of comparability because they are usually not given relative to the bank's size. The remaining indicators are standardized within regulatory frameworks and therefore comparable.

Only the Tier 1 capital ratio fulfills all four criteria. It is the most suitable quantitative risk indicator and will be used for the experiments within this thesis.

4.3.2 Discussion of Text Data Sources

In order to minimize noise and to enhance the sentiment analysis' validity, it is crucial to work with the documents well adapted to the task of risk sentiment analysis. Table 4.2 gives an overview whether each document type satisfies the four criteria described in Section 4.2, and the following paragraphs provide a discussion of each criterion.

Text Sources	Criteria: Suitable content	Accessible & available	Publication frequency	English language
CEO Letters	✓	✓	✓	✓
Risk Report	✗	✓	✓	✓
Outlook in Management Report	✓	✓	✓	✓
Press Releases	✗	✓	✓	✓
Internal CEO memos	✓	✗	✗	✗
Social Media	✗	✓	✓	✗

Table 4.2: Overview on the assessment of text data sources
Note: ✓ indicates *applicable*, ✗ stands for *not or just partially applicable*

Suitability. CEO letters, outlook sections, and internal CEO memos provide forward-looking and subjective information about attitudes and expectations towards risk. Risk reports are too objective, and press releases as well as social media channels do not provide suitable information for this thesis.

Accessibility and Availability. Besides internal CEO memos, all types of documents are publicly accessible and available—either via annual reports or other public channels.

Publication Frequency. Since internal CEO memos are not publicly available, it is unknown how often they are published. All of the other discussed document types are published in sufficiently short time intervals.

English Language. All of the analyzed banks are relatively large[11] and operate internationally. Hence, they publish annual reports and press releases not only in their local language, but also in English. However, internal CEO memos and social media posts are not necessarily written in English.

It can be concluded that the *Outlook section* in management reports and the *CEO letters* best fulfill the requirements. The latter are more subjective and might be better suited for qualitative analyses, e.g. how CEO language changes over time. Text from outlook sections is exclusively forward-looking and should therefore be well suited for predictive analyses.

[11] Otherwise, they would not be supervised via the SSM.

CHAPTER 5

Methodologies for Risk Sentiment Analysis

This chapter starts with explaining the process of data collection and documents the selection of a proper data mining tool. Another section is about data preprocessing and involves the extraction of relevant parts from the source documents as well as selection and representation of features. The structure of the first experiment is described in Section 5.3. It is about analyzing sentiment scores built up by incorporating the topic-specific word lists by Loughran and McDonald (2011). The objective of this experiment is to show how the language of forward-looking disclosures by European banks evolved within the last decade.

For the second experiment, the documents are labeled based on a quantitative risk measure, in particular the *Tier 1 capital ratio* (T1) dating to the end of the period referred to in the CEO letters and outlook sections. These data are then used for training supervised classification algorithms which aim to predict the indicator's evolution.

5.1 Preliminary Work

The previous chapter analyzed potential sources for text data and quantitative risk indicators. The next steps are to collect the data and to select a suitable tool for processing them.

5.1.1 Collection of Data

The following paragraphs describe how the data for the experiments were collected.

Annual Reports. The most suitable types of text data, i.e. CEO letters and the management report's outlook section, are published in the bank's annual reports. One way is to retrieve them directly from the banks' websites. All of the banks supervised via the SSM also provide an archive for downloading older annual reports. Unfortunately, these archives cannot be downloaded in a single pass, so one would rather have to download each document separately. In contrast to the United States, there is no central database for corporate filings in Europe[1]. Online company information systems like *Orbis* or *Bankscope* focus on quantitative data and do not provide batch downloads of periodical reports. *Bloomberg Terminals* provide batch downloads of periodic reports for individual banks, but it should be noted that this option is cumbersome[2]. Eventually, it was the best option to collect the annual reports for this work via a Bloomberg Terminal, supplemented by direct downloads from bank websites. In some cases, batch downloads of PDFs via a web browser add-on were possible. In total, 316 annual reports of 26 European banks were collected. All of them belong to the pool of 123 large banks which will be supervised via the SSM. The selected sample contains banks from all 12 countries which belong to the Eurozone at least since 2002. This sample promotes comparability of the data because the banks operated in similar economic circumstances and with the same currency. Each of the collected reports contains either a CEO letter, an outlook section, or both of these relevant sections.

Quantitative Risk Measures and Bank Data. The analysis in Section 4.1.2 showed that the Tier 1 Capital Ratio is the best suited quantitative risk indicator. Banks publish the respective figures in their periodic reports, but manually extracting them from the documents would be too cumbersome. Fortunately, there are two efficient ways for retrieving the figures.

The most important source for quantitative data is the online database *Bankscope*[3]. It provides information about banks from various countries, including the Eurozone, which is relevant for this work. The first step was to set the appropriate filters for the data. In particular, the data set was restricted to European banks, and the T1 was retrieved for the years between 2002 and 2013. The figures for 2014 will only be published in March or April 2015, so they cannot be included in this analysis. Downloading the data via Bankscope also delivers some additional information which is interesting for the evaluations, for example data about the bank's size.

An additional data source is the European Banking Authority. After the *stress test*, which was conducted in the course of establishing the Single Supervisory Mechanism, the EBA published detailed results for the participating 123 banks[4]. The data include the Tier 1 capital ratio as of December 2013 and whether the banks passed or failed the stress test.

Some figures were not contained in either of the data sets. In such cases, the numbers were manually extracted from the respective annual reports. In total, the following data points were collected and stored as CSV files:

- Country where the bank has its main residence.
- Bank name in a shortened and a full version.
- The bank's total assets, which represent its size.
- The bank's Tier 1 capital ratio at the end of each year between 2002 and 2013.

[1]In the US, annual reports and other filings by public companies are published via the EDGAR database. Many researchers in the field of financial text mining use this service for retrieving text data, see Section 3.4 of this work.

[2]At least for the purpose of this work, the downloaded files had to be individually reviewed in order to remove redundant files like slide decks and reports in other languages than English.

[3]See `https://bankscope.bvdinfo.com`, accessed September 21st, 2014.

[4]See `http://www.eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing/2014/results`, accessed December 14th, 2014.

Topic-Specific Word Lists. The lexicon-based approach in the first experiment requires word lists for categorizing the extracted features (cf. Section 3.3.4). The popular finance-specific word lists by Loughran and McDonald (2011) are freely available online[5] as CSV files. Besides the financial sentiment dictionaries, the authors also provide *stop word* lists.

5.1.2 Tools for Text Mining

A multitude of tools for text mining is available[6]. Many of them do not only support text mining, but are general data mining tools. Chen et al. (2007) evaluate 12 popular open source data mining systems. According to the authors, the following features are crucial for a good system:

- Ability to access and incorporate data from different sources and formats.
- Various functions for efficient data preprocessing.
- Availability of various data mining techniques since different problems require different solutions.
- Ability to cope with large data sets.
- Good visualization techniques for representing the data in an understandable way.
- Extensibility and interoperability in order to be open for new algorithms and other systems.
- An active community for updating and improving the system.

Comparison of Suitable Tools. The tools *GATE*[7] and *KNIME*[8] are of particular interest since they have been used at the *Information and Software Engineering Group* at Vienna University of Technology, where this work is written. Mikut and Reischl (2011) categorize data mining tools using nine different types. The authors assign the type *integration package* to KNIME since it belongs to the group of "extendable bundles of many different open-source algorithms" (Mikut and Reischl, 2011, p. 9). GATE focuses on a more narrow field, namely text mining, so it is categorized as *Solution*.

GATE is a popular system for natural language processing tasks. It is open software and contains—besides other modules—the information extraction system *ANNIE* and various machine learning algorithms. *GATE Developer* is a graphical interface which can be used for annotating documents and related tasks, which is the "basic business of GATE"[9].

[5]See `http://www3.nd.edu/~mcdonald/Word_Lists.html`, accessed December 14th, 2014.

[6]For a comprehensive list of text mining software, see `http://en.wikipedia.org/wiki/List_of_text_mining_software`, accessed December 23rd, 2014.

[7]GATE stands for *General Architecture for Text Engineering*, see Cunningham (2002).

[8]KNIME is the *Konstanz Information Miner*, see `http://www.knime.org/`, accessed December 30th, 2014.

[9]This quote is taken from `https://gate.ac.uk/sale/tao/splitch3.html#chap\protect\kern+.2222em\relaxdeveloper`, accessed December 29th, 2014.

While GATE would be the tool of choice for pure text mining tasks, it might be too inflexible with regards to the integration and analysis of quantitative data. This is where KNIME comes into play. It is a general purpose data mining tool, but features a strong text processing module introduced in Thiel and Berthold (2012). The latter provides various components for processing textual data. After transforming the data to a numeric representation, the regular KNIME data mining modules can be applied on the data as well. The KNIME analytics platform provides a visual workbench for the design of workflows via processing nodes and connections. The processing nodes provide methods for incorporating and preprocessing data, for calculations and statistical analyses, and for analyzing the data with various text and data mining algorithms. Like GATE, KNIME is also open source software and extensible. For example, it is possible to employ all of the WEKA[10] algorithms for the analyses. These capabilities and the more intuitive user interface make KNIME better suited for this thesis.

Text Processing in KNIME. The KNIME text processing plugin allows for reading, processing, and transforming textual data. This paragraph explains its main concepts and the underlying data structures based on Thiel and Berthold (2012).

KNIME offers several processing nodes for reading and parsing textual data, for example a node for flat text files as well as ones for parsing PDF, CSV, or DOC files. In any case, each parsed document is stored in a *DocumentCell*. This KNIME data type stores the text data and additional meta data like the file path or the number of terms contained in the document.

Furthermore, the text processing plugin provides nodes for tagging the parsed documents. Examples for this are a node for tagging parts-of-speech, a named entity tagger[11] for identifying persons or organizations, and a dictionary tagger for tagging terms according to given word lists.

For applying further preprocessing nodes, the *BoW Creator* node has to be applied on the document cells. It "transforms a data table consisting of a column containing *DocumentCells* into a bag of words, which is a data table consisting of a column containing *DocumentCells* and a column containing *TermCells*" (Thiel and Berthold, 2012, p. 7). Hence, each row of the resulting data table represents a specific term and a reference to the document in which it occurs.

The BoW Creator does not remove punctuation and numbers. If desired, they have to be deleted separately by subsequent processing nodes. Furthermore, KNIME provides linguistic preprocessing nodes which can be applied on the BoW data table. The available nodes include various filters, stemmers, and nodes for term replacements.

Finally, a range of nodes allows for the calculation of term frequencies. These figures are needed for term weighting approaches like the TF-IDF approach introduced in Section 3.3.2. KNIME stores the calculated figures in separate columns next to the preprocessed terms. Table 5.1 provides an excerpt of a KNIME data table in order to illustrate the representation of tagged terms, document references, and frequencies.

Regular data mining nodes in KNIME, e.g. for supervised classification, require a numeric representation of the data. This is done by a *Document Vector* node which creates a numeric

[10] The acronym WEKA stands for *Waikato Environment for Knowledge Analysis* and provides popular machine learning algorithms. See `http://www.cs.waikato.ac.nz/ml/weka/`, accessed December 30th, 2014.

[11] KNIME makes use of the *Apache OpenNLP* library, see `https://opennlp.apache.org/`, accessed February 13th, 2015.

Term	Document	TF abs	IDF
not[]	"chairman's statement...	2	0.348
obscure[]	"chairman's statement...	1	2.412
fact[]	"chairman's statement...	1	0.629
successful[POSITIVE(SENTIMENT)]	"chairman's statement...	3	0.488
years[]	"chairman's statement...	6	0.373
times[]	"chairman's statement...	1	0.633
challenging[NEGATIVE(SENTIMENT)]	"chairman's statement...	1	0.633
work[]	"chairman's statement...	2	0.463

Table 5.1: Excerpt of a KNIME data table which contains a BoW representation and related term frequencies.

vector for each document in the corpus. In particular, it creates a data table where each row represents a document vector[12]. The cells in this data table store either binary or numeric values. Binary values are used if only term presence / absence should be encoded, and numeric values store figures like TF-IDF weights.

Further details about the explained concepts and data structures are given in Thiel and Berthold (2012).

Set-up for the Experiments. The experiments are conducted with KNIME 2.10.4, its text processing plugin, and the WEKA 3.7 plugin. The software runs on a Intel Core 2 Duo CPU P8400 @ 2.26GHz with 4 GB RAM, the operating system is a 32-bit Ubuntu 14.04 LTS.

5.2 Reading, Preprocessing and Linking the Data

An important and usually laborious step of a text mining task is to prepare and preprocess the collected data. In this work, the first step is to extract the relevant sections from the annual reports and store them as plain text files. Afterwards, linguistic preprocessing is applied. In addition, this section also describes the process of reading and linking the collected data.

5.2.1 Building the Document Corpus

For building up the corpus, the original documents have to be parsed in order to acquire plain text files containing the required sections. Due to the heterogeneous structure of the annual reports, this is a challenging and time-consuming task. In the United States, 10-Ks have to follow a predefined structure. This is not the case in Europe, so the approach of Kogan et al. (2009) cannot be used[13]. Hence, the workflow depicted in Figure 5.1 had to be used for extracting CEO letters and outlook sections from the annual reports.

[12]The concept of document vectors is explained in Section 3.3.1.

[13]Kogan et al. (2009) could extract the *Management Discussion & Analysis* via a script.

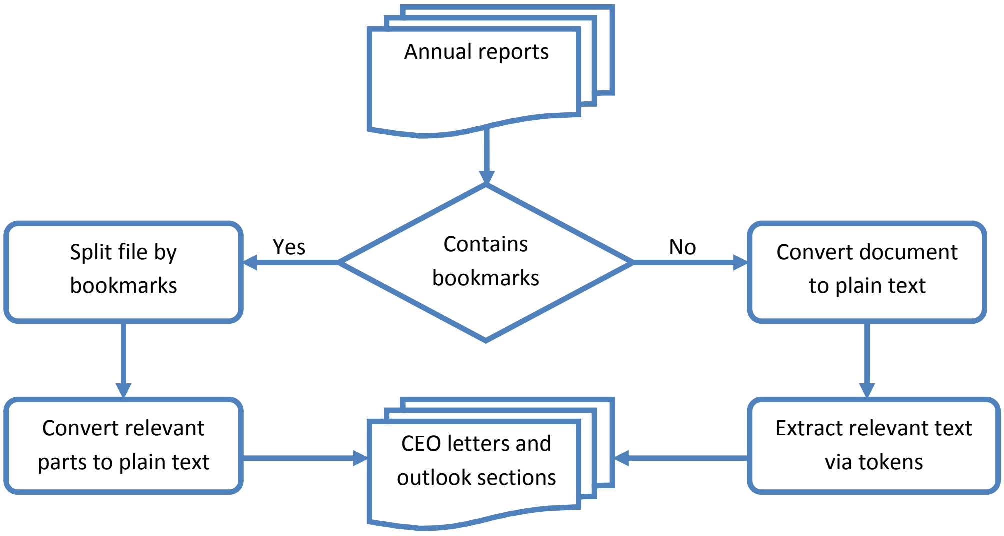

Figure 5.1: The workflow for extracting relevant text data from annual reports.

Converting PDFs to Plain Text. Although most text mining tools directly accept PDF files as a text source, it is useful to convert the files to plain text already during the process of corpus building. *PdfToText* is a popular command line tool for such tasks. The following command iterates through all PDF files in the current directory and converts them to text files. The numbers specify the crop area in order to exclude redundant information like headers and footers.

```
find '.' -maxdepth 1 -name '*.pdf' -type f -exec sh -c
  'pdftotext -x 40 -y 40 -W 516 -H 762 "$0"' {} \;
```

Split Original Documents. If the original PDF file contains bookmarks for navigating within the document, it is possible to split the file corresponding to the sections of the document. In a next step, the relevant sections are selected and converted into plain text. Unfortunately, such bookmarks are only present in a small share of reports and often not sufficiently fine-grained, e.g. if the bookmarks only refer to the management report but not to the outlook section within it. For realizing the document splitting, the Linux tool *Sejda* was used. The following command iterates through all PDF files in the current directory and its subdirectories. If a PDF file contains bookmarks, it is split according to these. The caption of the split files is defined by the original file's name and the bookmark name.

```
find '.' -maxdepth 2 -name '*.pdf' -type f -exec sh -c
  'sejda-console splitbybookmarks -l 2 -f "$0" -o .
  -p [BASENAME]_[BOOKMARK_NAME]' {} \;
```

Extracting Relevant Text. If the annual report does not contain meaningful bookmarks, the PDF files are first converted to plain text. The next step is to extract the relevant sections which are delimited by specific *tokens*. For example, a typical CEO letter is delimited by the tokens *Dear Shareholders* and *Sincerely*. The following bash script searches the converted text files for the specified tokens and stores the text between the tokens in a separate file. This technique was used for the majority of the extracted CEO letters and outlook sections.

```
#!/bin/bash
TOKEN_BEGIN="Dear Shareholders,"
TOKEN_END="Sincerely,"

mkdir -p ./extracted_letters

for FILE in *.txt
do
        BASE=$(basename "$FILE" .txt)
        sed -n -r "/$TOKEN_BEGIN/,/$TOKEN_END/w
          ./extracted_letters/${BASE}_ceo_letter.txt" $FILE
done
```

In order to make the extracted text files uniquely identifiable, a standardized naming convention has been applied. The text files follow the convention *name-of-bank_year_type-of-publication.txt*. The documents were published between March and May of the year given in the filename, obviously corresponding to the publication date of the annual report. The type of publication is either *ceo_letter* or *outlook*.

Document Corpus Overview. With the preprocessing steps described in this section, separate plain text files were generated. Each of the files contains either a CEO letter or an outlook section. In total, 530 documents were extracted from the banks' annual reports. The corpus is composed by documents published between 2002 and 2014 and incorporates banks from each country that has been belonging to the Eurozone at least since 2001. Table 5.2.1 gives an overview of the number of documents in the corpus. It shows that the number of published outlook sections constantly increased between 2002 and 2014. The number of CEO letters also increased over the years, but only until 2008, when some CEOs stopped to write letters in the course of the financial crisis[14] .

5.2.2 Reading and Linking the Data

The first task of the data analysis with KNIME is to read and parse the text files as well as the additional quantitative data. Afterwards, these data have to be joined. The respective KNIME workflow is depicted in Figure 5.2 and explained in the following paragraphs.

[14]It should be mentioned that 6 documents labeled "Letter to the Shareholders" from the Italian bank MPS were not included in the analysis. The reason is that they are rather complete management reports than CEO letters. Including them in the analysis would bias the results.

Year	# of CEO letters	# of outlook sections
2002	15	14
2003	19	19
2004	17	20
2005	19	20
2006	19	21
2007	23	22
2008	25	23
2009	21	23
2010	20	23
2011	21	23
2012	20	23
2013	22	24
2014	22	23
Total	**263**	**278**

Table 5.2: An overview of the document corpus.

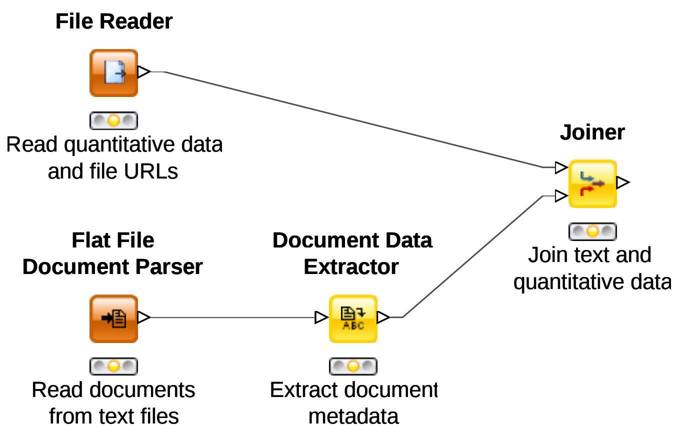

Figure 5.2: The workflow for reading, parsing and linking the collected data in KNIME.

The KNIME text processing module provides several nodes for importing data from files. The outlook sections respectively the CEO letters were read via a *Flat File Document Parser*, which imports all text files in a predefined directory and stores them in KNIME *DocumentCells*. The subsequent node extracts meta data and stores them alongside to the document. These data involve the document category (either CEO letter or outlook), the file path, and the contained number of terms.

A *File Reader* node parses the CSV file with the quantitative indicators and the additional information as described in Section 5.1.1. It stores the data in an internal table and converts the contained quantitative figures to the number formats *integer* or *double*.

Finally, the extracted data need to be linked in order to make them ready for the actual experiments. The standardized and unique file name serves as join attribute. The *Joiner* node is configured to perform a *right outer join*, i.e. it includes *all* text documents and links the quantitative data to them. If the latter are not available, null values are written in the respective cells. In fact, this is only an issue for the text data published in 2014, where the corresponding quantitative data have not been available when this thesis was written.

5.2.3 Linguistic Preprocessing

This subworkflow removes irrelevant features from the text and prepares the remaining ones for subsequent steps in the workflow. The respective KNIME processing nodes are organized within a meta node[15] depicted in Figure 5.3. The preprocessing nodes operate on the *bag of words* (BoW) representation of the documents explained in Section 5.1.2.

Description of the Individual Linguistic Preprocessing Nodes. The first node removes punctuation characters in terms since they are not relevant for this type of analysis. The same holds for numbers, so all rows where the extracted term consists only of numbers are removed in the next step. The third node is a *N Chars Filter*. It is configured to remove all rows where the term consists of less than 2 characters. Hence, it removes short words like *a* or *I* as well as single characters, for example ones which had been used in a list of items as *a)*, *b)*, and *c)*. The next node is a *stop word filter*. It filters rows with common words which have no relevance in the context of risk sentiment analysis. Instead of the built-in stop word list provided by KNIME, lists provided by Loughran and McDonald are used for this filter[16]. The reason for this is that the built-in list contains words which carry sentiment in a financial context, for example *almost* or *problem*. Finally, a *Case converter* node converts all items in the BoW to lower case terms. This is important for the frequency measures, where it must not make any difference whether terms are written in lower or upper case.

[15] KNIME meta nodes appear as single node in the workflow, but incorporate several processing nodes. Meta nodes can be reused and make it easier to structure the workflow. See `https://tech.knime.org/metanodes`, accessed January 3rd, 2015.

[16] The stop word lists by Loughran and McDonald are provided at `http://www3.nd.edu/~mcdonald/Word_Lists.html`, accessed January 3rd, 2015. In particular, the lists *Generic* and *Dates and numbers* were incorporated.

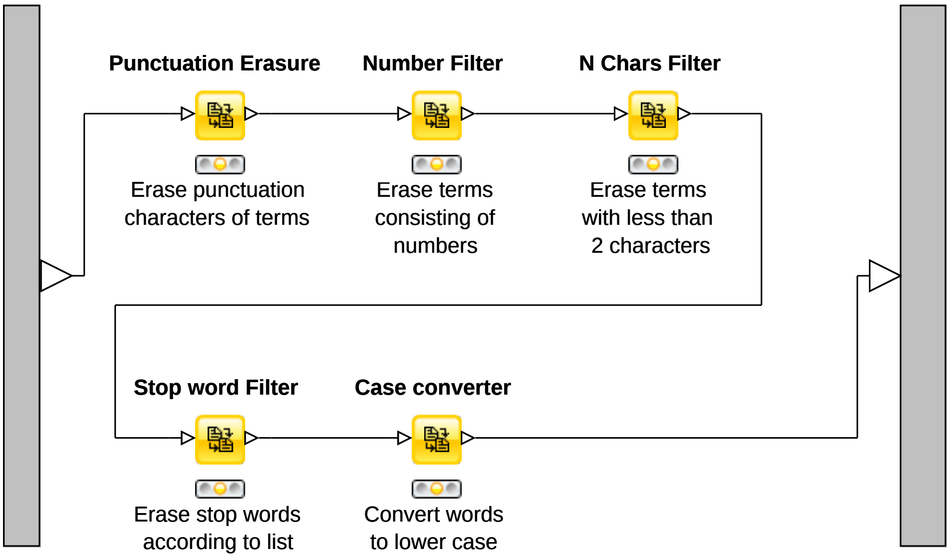

Figure 5.3: The workflow for linguistic preprocessing in KNIME.

Further Remarks. It should be mentioned that some popular preprocessing steps were omitted in this workflow. For example, deletion of markups is not necessary because the document corpus already consists of clean plain text files (see Section 5.2.1). Furthermore, *stemming* can be omitted as well—at least for the lexicon analysis—because the word lists provided by Loughran and McDonald (2011) also include inflected versions of the sentiment words. According to Loughran and McDonald, stemming faces the problem that adding common prefixes or suffixes could change a word's meaning. Hence, the authors included only such inflections which retain the root word's meaning.

5.3 Lexicon-based Analysis

This experiment analyses the text data with proven methods of sentiment analysis in order to prepare them for the qualitative evaluation. The workflow roughly follows Kearney and Liu (2014) and is presented in Figure 5.4. It consists of the following steps: (1) reading, parsing, and linking the collected data, (2) the actual sentiment analysis which derives the scores, (3) data consolidation, and (4) data evaluation. Step 1 has already been explained in Section 5.2.2, so this section starts off with step 2.

5.3.1 Calculating Sentiment Scores

This part of the workflow receives the parsed text data as input and performs sentiment tagging, linguistic preprocessing, valence shifting, term weighting, and finally calculates sentiment scores for each document. The following paragraphs describe only those steps which have not

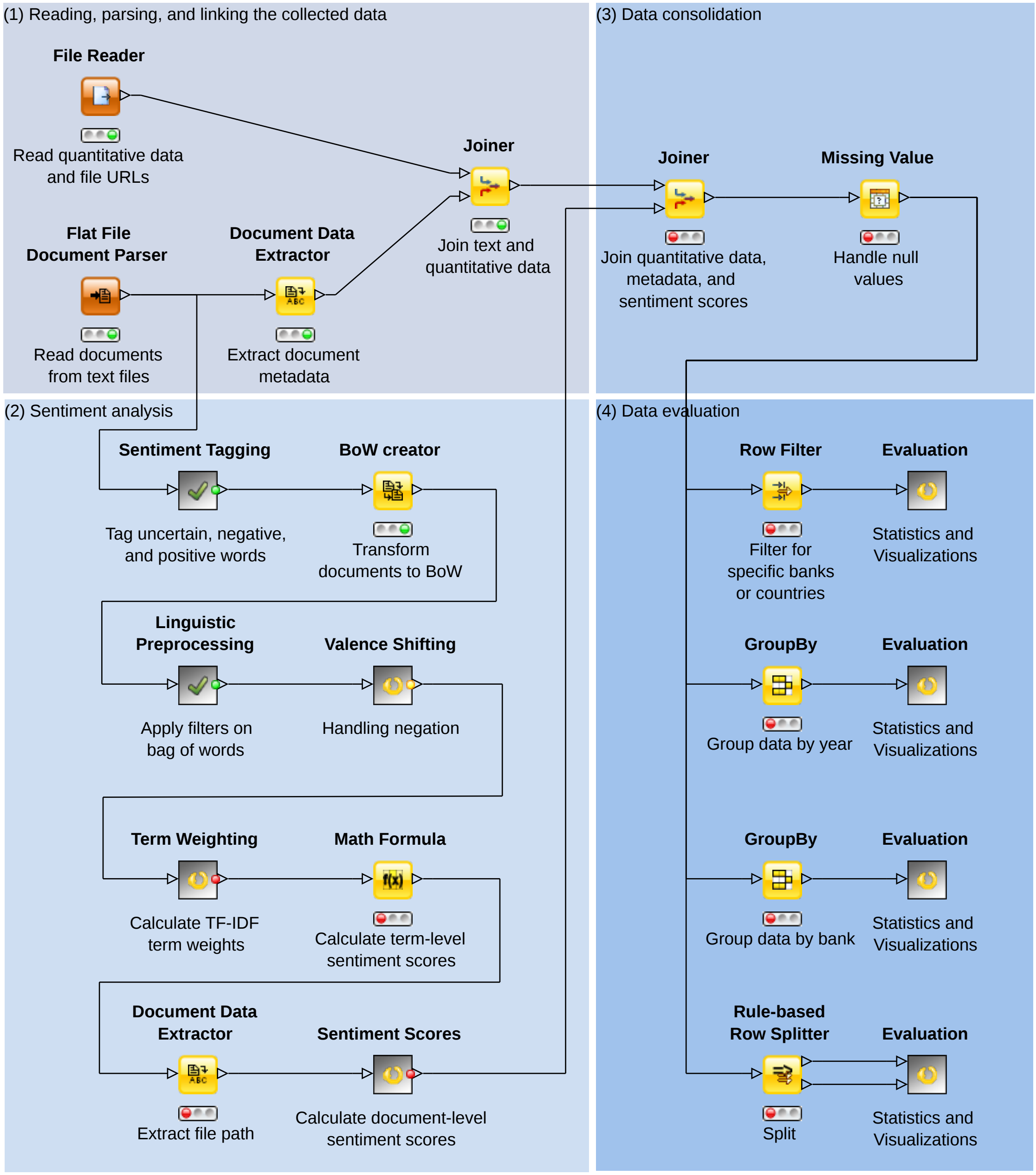

Figure 5.4: The workflow for lexicon-based sentiment analysis in KNIME.

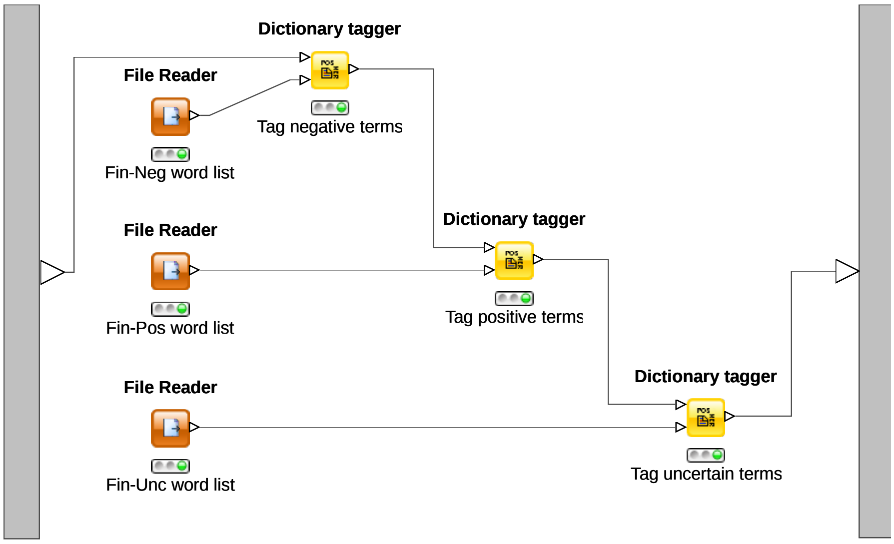

Figure 5.5: The workflow for sentiment tagging in KNIME.

yet been discussed, i.e. it does not include the BoW creation and the linguistic preprocessing steps.

Sentiment Tagging. Like the linguistic preprocessing, sentiment tagging is also contained in a meta node. The content of this meta node is depicted in Figure 5.5. Three file readers parse the finance-specific word lists provided by Loughran and McDonald (2011). In particular, this thesis works with negative words (*Fin-Neg*), positive words (*Fin-Pos*), and words related to uncertainty (*Fin-Unc*). The lexicons are stored in tables which serve as inputs for the *dictionary taggers*. The latter assign a tag to a term in the document if it appears in the word lists. The tagging is configured to require exact matches and to work insensitive to case. Finally, all positive, negative, and uncertain words are assigned with the respective tag. The following paragraph illustrates this with a tagged excerpt of a CEO letter. Uncertain terms are highlighted in yellow, negative ones in red, and positive ones in green.

> “We remain cautiously optimistic about the years to come. Generally speaking, the macroeconomic climate in our home markets is encouraging . In saying that, of course, we are also fully aware of the challenges ahead of us. The budgetary problems confronting certain European countries continue to be a source of un-certainty ” (KBC Group NV, 2011, p. 9).

Valence Shifting. The processing nodes depicted in Figure 5.6 implement the method proposed by Polanyi and Zaenen (2006) for handling negation. The first processing node adds three columns to the data table. These columns contain the three direct predecessors of each term.

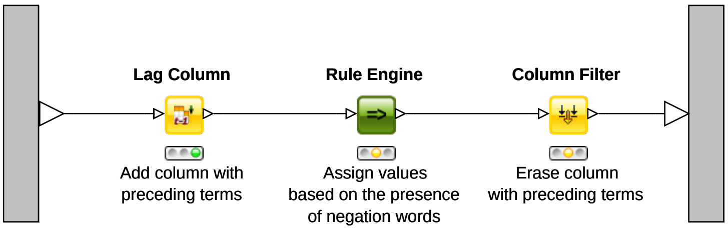

Figure 5.6: The workflow for valence shifting in KNIME.

Technically, this is achieved by copying and shifting the term column[17]. Another new column stores whether one of the three direct predecessors contains a negation word or not. Based on Polanyi and Zaenen (2006) and Loughran and McDonald (2011), the following negation words are used: *no*, *not*, *don't*, *never*, *none*, *neither*. The last node deletes the shifted term columns since they are not needed anymore.

Term	Term(-1)	Term(-2)	Term(-3)
not[]	sector[]	banking[]	european[]
running[]	not[]	sector[]	banking[]
smoothly[POSITIVE(SENTIMENT)]	running[]	not[]	sector[]

Table 5.3: Excerpt of a KNIME data table showing the original and shifted term columns.

Consider the text snipped *not running smoothly* as an example. Table 5.3 shows the related excerpt of the KNIME data table with the original and the shifted term columns. The term *smoothly* is tagged as positive, but the preceding terms include a negation word. Hence, the term's sentiment score will be negative.

Term Weighting. The outcome of this meta node is a weight for each term, calculated according to Equation 3.2 in Section 3.3.2. The corresponding workflow is shown in Figure 5.7. It starts with an IDF node provided by KNIME, which calculates the *inverse document frequency* for each term. The next node computes each term's *absolute frequency* in the respective documents. Since these absolute frequencies should be represented in a logarithmized form, a *Math Formula* node processes the corresponding calculation. The remaining nodes deal with document length normalization and calculate the final term weights.

Calculating Sentiment Scores. A multiplication of the already computed TF-IDF weight by the valence shifter variable results in the term-level sentiment score. These figures serve as a basis for calculating document-level scores for uncertainty, positivity, and negativity. The corresponding workflow is depicted in Figure 5.8.

[17]The BoW creator node adds new rows to the data table according to the original document's word order. Hence, copying and shifting the term column makes it possible to identify the predecessors.

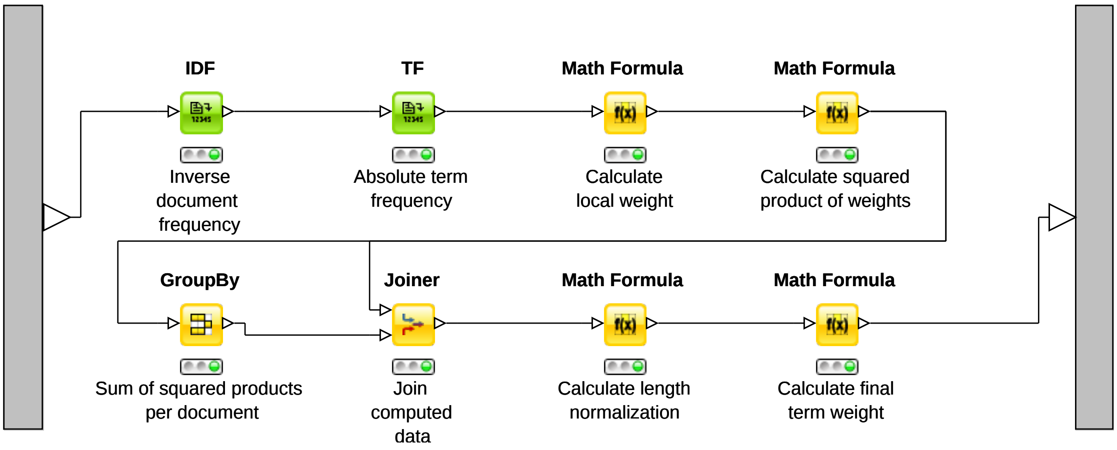

Figure 5.7: The workflow for term weighting in KNIME.

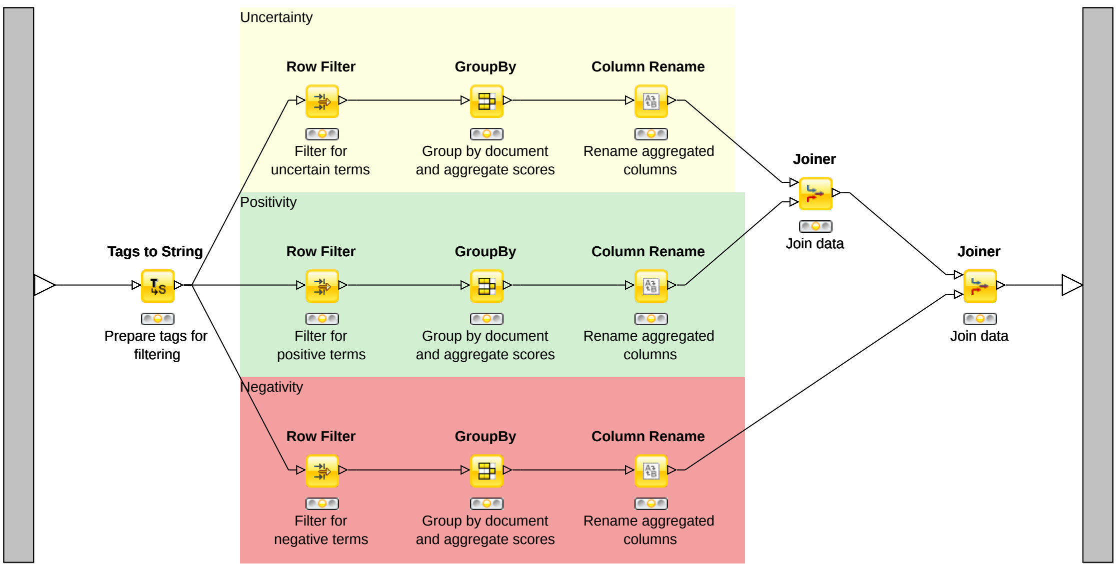

Figure 5.8: The workflow for calculating document-level sentiment scores in KNIME.

First, the tags are extracted as strings in order to prepare them for the subsequent *Row Filter* node. Afterwards, three analog sub-workflows conduct the selection of features, i.e. they filter all terms besides the ones with the sentiment tags. The next nodes group the table by the documents and sum up the sentiment scores from the individual features. The third group of nodes rename the score column to allow for correct identification in the subsequent steps. Finally, the columns with the sentiment scores are joined. After these steps, a sentiment score for each of the three categories is available for every analyzed document. Hence, the document sentiment score $s_{c,j}$ for a specific sentiment category c and a document j is computed according to:

$$s_{c,j} = \sum_{i=0}^{n} w_{i,j} v_i \tag{5.1}$$

The document sentiment score is the sum of the n term sentiment scores which belong to the document j and the sentiment category c. The latter is either negativity, positivity, or uncertainty. The term sentiment scores are computed by multiplying the term weight $w_{i,j}$ by the valence shifter v_i.

5.3.2 Data Consolidation and Visualization

Steps 3 and 4 of the lexicon-based analysis workflow in Figure 5.4 consolidate the data, handle missing values and prepare the data for the evaluations discussed in the next chapter. First, the quantitative data extracted in step 1 are joined with the sentiment scores from step 2. This is done via a *left outer join* in order to keep all documents in the table.

Missing Values. It is possible that some sentiment score cells are empty. This is the case when no sentiment words of a category were present in the document. For example, if a CEO letter does not contain any words which appear in the Fin-Neg word list, its negativity score would be null[18]. In such a case, it is safe to overwrite this null value with zero. Another potential reason for null values are missing documents. It this case, it would be misleading to overwrite the null values. However, since the left outer join ensures that there are no missing documents in the resulting table, this is not an issue. The node *Missing Value* handles the null values by replacing them with zero.

Grouping and Representing the Data. The last part of the workflow groups and represents the data for evaluation purposes. In particular, the following nodes have been integrated in the workflow:

- A *Row Filter* for quick extraction of data for a specific bank, a specific country, or a specific bank size.

[18] An example is the 2005 annual report of *Bank Austria* which was published in March 2006. It consists of 1075 terms, only two of which are uncertain, no ones are negative, and a stunning 41 are positive. The CEO letter starts with: "Ladies and Gentlemen, In 2005, Bank Austria Creditanstalt achieved the best results in the bank's history" (Bank Austria Creditanstalt, 2006, p. 7).

- Two *GroupBy* nodes which group the data by year respectively by bank. These data are important for analyzing the evolution over time and across the different banks.
- A *Rule-based Row Splitter* divides the data set into two parts based on groups of countries in distinct economic situations.

The KNIME modules *Data Views* and *Statistics* provide many processing nodes for representing and visualizing data. The most useful visualizations for the evaluations are bar charts, line charts, and scatter plots. Furthermore, special visualizations like *tag clouds* give a quick overview on the most important terms.

5.3.3 Statistical Analysis

In addition to a qualitative analysis of the sentiment scores, the evaluation of the lexicon-based approach will also make use of some statistics, in particular *correlation coefficients* and *linear regressions*.

Correlation Coefficients. The best way to calculate a correlation coefficient r in KNIME is to use a *Linear Correlation* processing node, which determines r for each pair of columns chosen in the node. Formally, r is calculated by

$$r(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{5.2}$$

where $r(X, Y)$ denotes the correlation coefficient between the data sets X and Y. The numerator σ_{XY} denotes the *covariance* between X and Y, which is a measure for the linear dependency between the data sets. In the denominator, σ_X and σ_Y are the *standard deviations* of the data sets.

For the correlation coefficient, it holds that $-1 \leq r \leq 1$. The value $r = 1$ would indicate a perfect positive linear correlation, $r = 0$ no correlation, and $r = -1$ a perfect negative linear correlation (cf. Wooldridge, 2002, p. 680ff).

Multiple Linear Regression Analysis. A fundamental method for explaining a dependent variable y by means of one or more independent variables x_i is the *multiple linear regression analysis*. A regression model with k independent respectively explanatory variables is given by

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \tag{5.3}$$

where the dependent variable y is explained by the following terms: the *intercept* β_0 denotes the value of y when all the independent variables x_i are zero, i.e. $x_1 = \cdots = x_k = 0$. The *ceteris paribus effect*[19] of x_i on y is measured by the coefficient β_i. The error term u collectively contains all factors affecting y which are not covered by the independent variables (cf. Wooldridge, 2002, p. 69ff).

[19] In this context, ceteris paribus means that the effect measured by β_i only holds when all other factors affecting y are fixed.

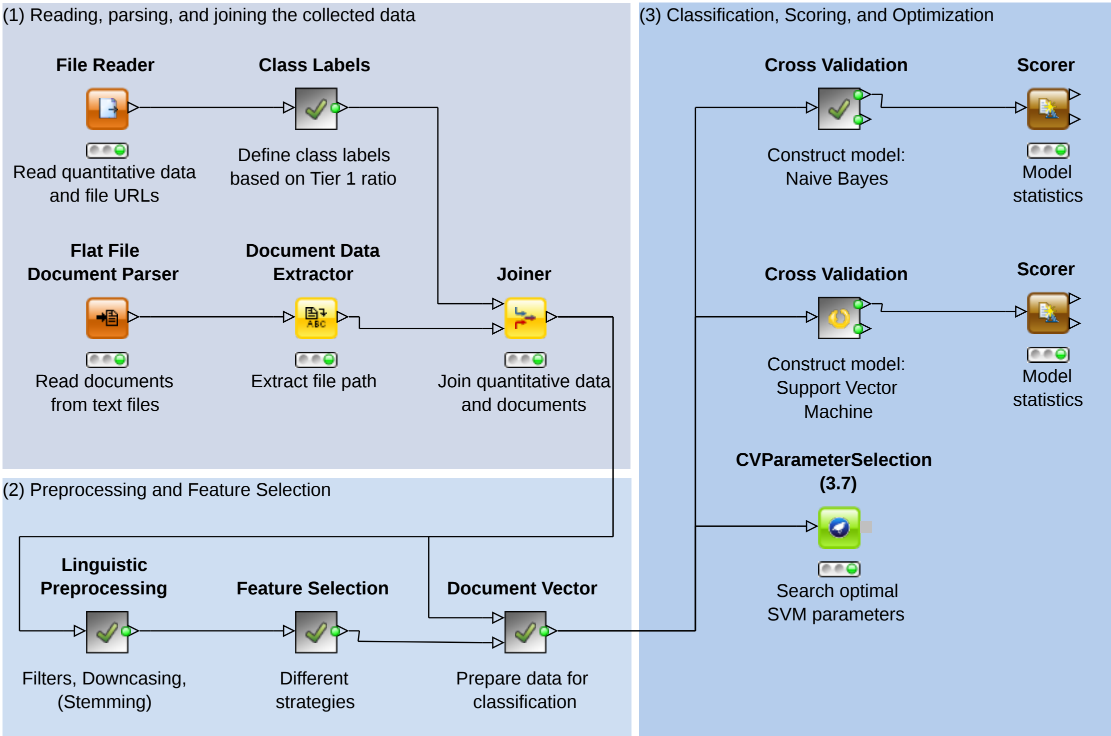

Figure 5.9: The workflow for supervised classification in KNIME.

The coefficients β_i are approximated by the *ordinary least squares* (OLS) method, which aims to minimize the sum of squared residuals[20]. A detailed description of OLS is given in Wooldridge (2002, p. 27ff).

The multiple linear regression analysis depends on a number of assumptions which might not hold for real-world problems. For example, it "requires that all factors in the unobserved error term be uncorrelated with the explanatory variables" (Wooldridge, 2002, p. 70). Nevertheless, it is "still the most widely used vehicle for empirical analysis in economics and other social sciences" (Wooldridge, 2002, p. 66).

5.4 Analysis based on Supervised Classification

The objective of this experiment is to test whether proven methods of *supervised classification* are suitable for predicting the evolution of quantitative risk indicators, in particular the T1. The general workflow is presented in Figure 5.9. It consists of three steps: (1) reading and parsing the collected data as well as assigning the class labels, (2) linguistic preprocessing and two

[20] A residual is the vertical distance between an individual data point and the linear slope given by y.

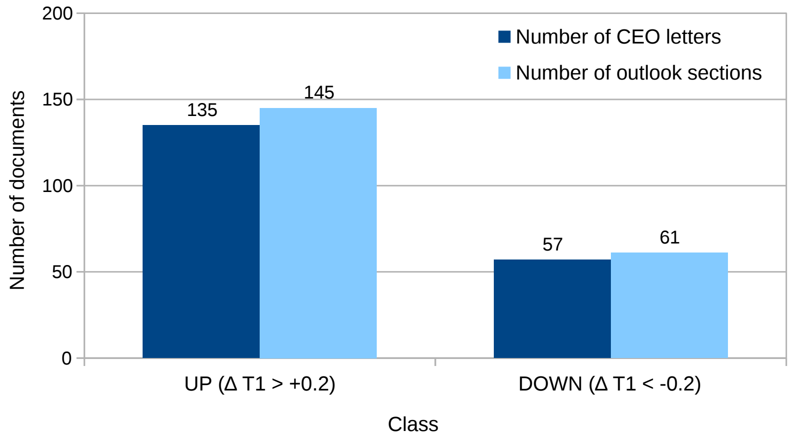

Figure 5.10: The distribution of class labels for CEO letters and outlook sections.

different feature selection strategies[21], and (3) classifying the data with *Naïve Bayes* (NB) and *Support Vector Machine* (SVM). In addition, the last step contains *Scorer* nodes for measuring the performance of the classification algorithms as well as a node for parameter optimization.

5.4.1 Assigning the Class Labels

Reading, parsing, and linking the data has already been explained in Section 5.2.2, so it is sufficient to explain the method for assigning the class labels in this chapter in order to cover step 1.

The Tier 1 capital ratio is published by banks at least once a year. Since it is actually a continuous measure, it always strongly depends on the previous year's ratio. Banking supervisors like the ECB are interested in the future evolution of the ratio: if it increases, the bank acts in a less risky way[22], and vice versa. Hence, appropriate labels for the supervised classification task are *UP* for an increasing T1, and *DOWN* for a decreasing one. In addition, one could assign the label *EQUAL* if the difference of the ratios is below a specific threshold. Unfortunately, preliminary tests showed that such a *multinomial classification* does not perform in a satisfactory way, so this option cannot be considered.

Figure 5.10 shows the class distribution for the chosen binary classification. In order to achieve a better separation of the classes, the thresholds are defined as follows: if the Tier 1 capital ratio increased by at least 0.2 pp[23] compared to the previous year, the class label *UP* is assigned. On the other hand, if the T1 *decreased* by more than 0.2 pp, the assigned class label is *DOWN*. This method is also referred to as *binning*, i.e. the discretization of numeric data.

[21] Two versions of the meta node *Feature Selection* are used. One implements a feature selection strategy based on word lists, and the other one makes use of *document frequency* and *information gain* (see Section 5.4.2).

[22] It acts in a less risky way because (1) the "good" Tier 1 capital increased, (2) there are less risk-weighted assets, or (3) a combination of (1) and (2).

[23] The unit *pp* refers to *percentage points*, which measure differences between percentages. Hence, an increase of 0.2 pp indicates a change of the Tier 1 capital ratio from e.g. 10 % to 10.2 %.

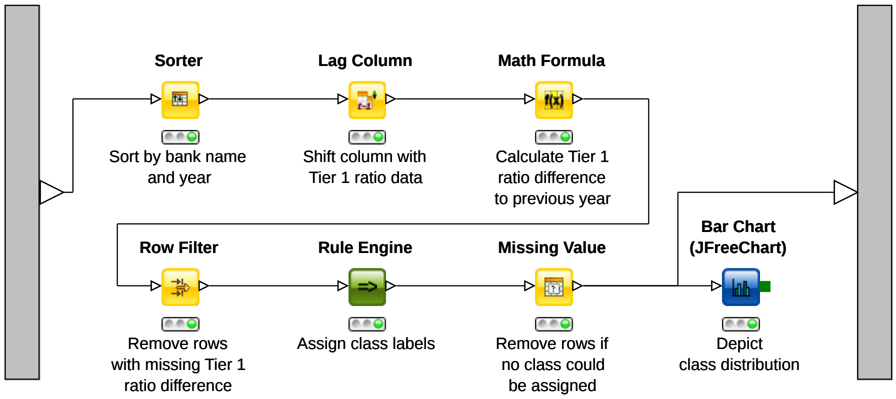

Figure 5.11: The workflow for assigning class labels in KNIME.

Figure 5.11 depicts the workflow for assigning class labels in KNIME. First, the data are sorted in order to compare the correct values when the column with the T1 is shifted by one cell in the next step. The *Math Formula* node calculates the difference between each Tier 1 capital ratio and the corresponding figure from the previous year. After removing the instances where no difference could be calculated[24], the class labels are assigned by a rule engine. Finally, instances where the T1 did not change notably[25] are removed. This decreases the number of available instances for training and testing the classification algorithms, but the remaining ones can clearly be assigned to either *UP* or *DOWN*.

5.4.2 Preprocessing and Feature Selection

Linguistic preprocessing comprises removing punctuation, numbers, single characters, and stop words. The remaining words are converted to lower case. All these steps are explained in Section 5.2.3. Furthermore, the terms are weighted according to the TF-IDF strategy presented in Section 3.3.2. Preliminary tests showed that TF-IDF yields slightly better results than a simple *term presence* approach. For feature selection, two approaches are followed.

Feature Selection Based on Sentiment Words. This approach assumes that the sentiment words used in the lexicon-based analysis are the relevant features for this experiment. Hence, all the words which do not appear in the word lists by Loughran and McDonald (2011)[26] are removed. Term weighting respectively calculating the sentiment scores does not differ from the lexicon-based analysis.

Feature Selection With Document Frequency and Information Gain. It is unclear whether the terms in the word lists are relevant for the classification purpose of this experiment. Hence,

[24] The difference between 2013 and 2014 could not be calculated because the T1 for 2014 are not yet available.
[25] This is the case when the T1 difference is between -0.2 pp and +0.2 pp.
[26] Like in the lexicon-based analysis, the lists *Fin-Neg*, *Fin-Pos*, and *Fin-Unc* are used.

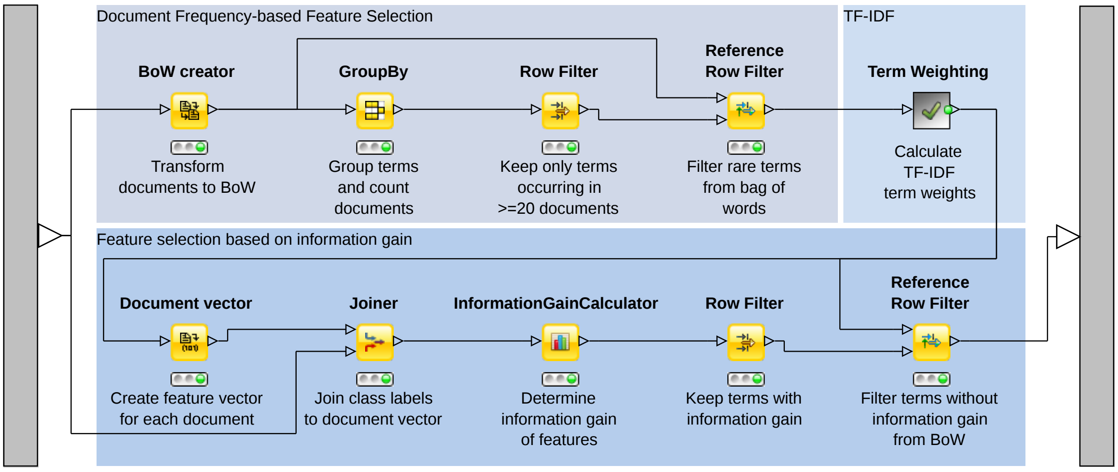

Figure 5.12: The workflow for feature selection based on DF and IG in KNIME.

the second feature selection approach follows a completely different strategy.

Besides an additional *Snowball Stemmer*, linguistic preprocessing does not differ from the previous approach. The stemmer ensures that different versions of the same word are treated as equal. The KNIME workflow for this feature selection approach is depicted in Figure 5.12. It implements a strategy based on *document frequency* (DF) and *information gain* (IG)[27]. The aim of DF is to consider only terms which appear in at least 20 distinct documents[28]. In the KNIME workflow, this is achieved by grouping terms and counting the corresponding documents. This threshold was selected because it led to the best results in preliminary tests[29]. Afterwards, the rare terms are filtered from the bag of words. Based on the resulting BoW, the term weights are calculated.

In order to calculate the information gain for each feature, the respective KNIME processing node requires document vectors[30] with associated class labels. After creating the vector and calculating the IG figures, a *Row Filter* removes all terms with an information gain close to zero from the BoW representation.

5.4.3 Classification, Scoring, and Optimization

The outcome of steps 1 and 2 in the workflow (see Figure 5.9) is a set of document vectors with associated class labels. These data can be analyzed by a multitude of classification algorithms. This thesis employs the widespread methods *Naïve Bayes* (NB) and *Support Vector Machine*

[27]Both methods are explained in Section 3.3.3.

[28]This corresponds to 10.5 % of the CEO letters respectively 9.7 % of the outlook sections used in the supervised classification task.

[29]The tested thresholds were 1, 20, 40, 60, and 80 documents. 20 yielded the best classification accuracy.

[30]KNIME creates a vector for each document. The vector's dimension equals the number of distinct terms in the bag of words.

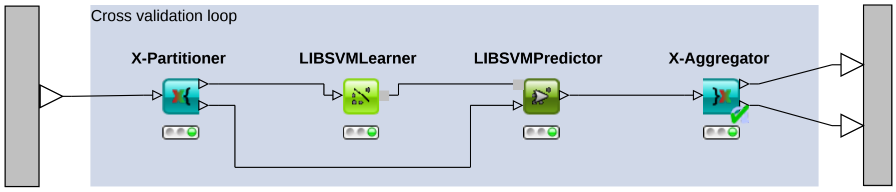

Figure 5.13: The cross validation meta node in KNIME.

(SVM). The latter is used in its basic version, i.e. with a linear kernel. This section presents the classification and evaluation concepts utilized in the second experiment.

Cross Validation. Training and testing a supervised classification algorithm requires distinct data sets. Otherwise, the test results would be biased. A crucial decision is how to divide the document corpus in a training and a test set. In general, more training data results in a better classifier. If the share of documents used for training purposes is too small, the classifier would be overfitted. Such a classifier works well with the training set, but has a bad performance if unseen instances should be classified, i.e. it cannot generalize. Another issue arises if the model is validated by only one test set, since it might work well on that particular set but not in general.

A popular approach for tackling these problems is *k-fold cross validation*. This approach splits the document corpus into k disjoint parts of equal size. The classifier is then trained with $k-1$ parts and tested on the remaining part. This procedure is repeated k times until each of the sets was used once for testing. The classifier's overall accuracy is then estimated by calculating the mean of the individual classifiers' accuracies. A value of $k = 10$ is popular, as it allows for using 90 % of the data as the training set (Hsu et al., 2003; Gupta, 2006, p. 130f).

In KNIME, cross validation is handled by a special meta node. Figure 5.13 depicts the processing nodes within this meta node. A *X-Partitioner* and a *X-Aggregator* handle the k iterations. The former splits the data set, the latter collects the results. Within each iteration, a learning node trains a specific classifier, in this case a LIBSVM learner. The subsequent node predicts the classes of the test set's instances.

Optimization of Parameters. While Naïve Bayes works without parameters, all variants of SVMs depend on at least one parameter. The only parameter for a linear SVM is C, which "controls the tradeoff between margin maximization and error minimization" (Chapelle et al., 2002, p. 133). In case C is too small, the classifier will be inaccurate, i.e. it is not able to classify the instances in a satisfactory manner. On the other hand, a too large C would lead to an overfitted SVM.

Hsu et al. (2003) recommend to optimize SVM parameters by conducting a *grid search*. This simple method trains and evaluates the SVM with several parameter values within a predetermined range. It is useful to start with a coarse grid for identifying C's order of magnitude. In a next step, a good value for C can be determined via tests on a finer grid. The WEKA plugin in KNIME provides the node *CVParameterSelection* for this purpose. This processing node performs a parameter selection by cross validation between a given lower and upper bound. Furthermore, the number of optimization steps between these bounds can be specified.

Performance Scores. After training the classifiers by utilizing cross validation, a *Scorer* node calculates measures for assessing the performance of each algorithm. Among the established performance measures for supervised classification are *accuracy*, *precision*, and *recall*. It is also common to present the results by means of a *confusion matrix*. The following paragraphs explain these concepts based on Gupta (2006) and Fawcett (2006).

In general, "the accuracy of a classification method is the ability of the method to correctly determine the class of a randomly selected data instance" (Gupta, 2006, p. 128). The accuracy measure in KNIME determines the percentage of correctly classified instances.

		True class	
		A	Not A
Predicted class	A	True Positives (TP)	False Positives (FP)
	Not A	False Negatives (FN)	True Negatives (TN)

Table 5.4: Interpretation of a confusion matrix.

Table 5.4 illustrates the interpretation of a confusion matrix for a binary classification problem. Its elements can be explained as follows: if the classifier predicts class A and the instance indeed belongs to class A, the result is a *true positive* (TP). On the other hand, if the classifier predicts class A, but the true class is not A, the result is a *false positive* (FP). *False negatives* (FN) and *True negatives* (TN) can be defined in an analog way.

These definitions serve as a basis for the other measures. Precision is defined as the share of instances classified as positive which are indeed positive:

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

Recall is the quotient between correctly classified positives and the total number of positives[31]:

$$Recall = \frac{TP}{TP + FN} \tag{5.5}$$

Which of these performance measures is the most important one? There is no clear answer to this question. If the set of instances classified as A should include all of the true A's, one should maximize recall. However, this could lead to a large number of false positives. If false positives should be avoided, it is better to choose a classifier with high precision.

[31] In this work, precision and recall are explained in the context of classification. In *information retrieval*, precision would be the share of retrieved documents which are indeed relevant. Recall would be quotient between the retrieved relevant documents and the total number of relevant documents.

CHAPTER 6

Evaluation of the Experiments

In this chapter, conclusions are drawn from the results of the experiments. The figures from the lexicon-based approach are analyzed in order to explain how risk sentiment of banks in the Eurozone have been evolving since 2002. Furthermore, different groups of countries are compared. The supervised classification experiment tries to predict the evolution of the Tier 1 capital ratio (T1) based on bank disclosures. The evaluation of this approach discusses the performance of the different algorithms, and tests whether it is better to use CEO letters or outlook sections for the predictions.

6.1 Evaluation of the Lexicon-based Approach

The outcome of the lexicon-based approach consists of sentiment scores for each document representing the degrees of uncertainty, negativity, and positivity. All of them are interesting with regards to risk sentiment analysis: uncertainty relates to risk in a direct way. Highly negative sentiment refers to current or future problems, and too positive sentiment could represent overconfidence. Evaluating these sentiment scores with classical performance measures would require the manual classification of a training set with labels like *certain* versus *uncertain*. Regarding the aims of this thesis, it is much more interesting to conduct a qualitative analysis, i.e. to put the figures in a historical context and to analyze how the financial crisis and other events were reflected in the bank disclosures. Nevertheless, the sentiment scores are also compared to quantitative figures like bank size and the Tier 1 capital ratio.

The tag cloud in Figure 6.1 gives a first overview about the sentiment words in the analyzed CEO letters. The more often a word was used in these documents, the larger it is represented in the tag cloud. The colors encode the sentiment categories as explained in the picture's caption. It is not surprising that words like *risk*, *crisis*, and *difficult* belong to the most frequent sentiment words from the Fin-Unc and Fin-Neg word lists. Often used positive words are *strong*, *good*, and *achieved*. It is worth to mention that these frequently used words did not necessarily receive a high sentiment score because of the TF-IDF term weighting scheme.

Figure 6.1: A KNIME tag cloud with the most frequent sentiment words. Yellow words refer to uncertainty, green to positivity, and red to negativity.

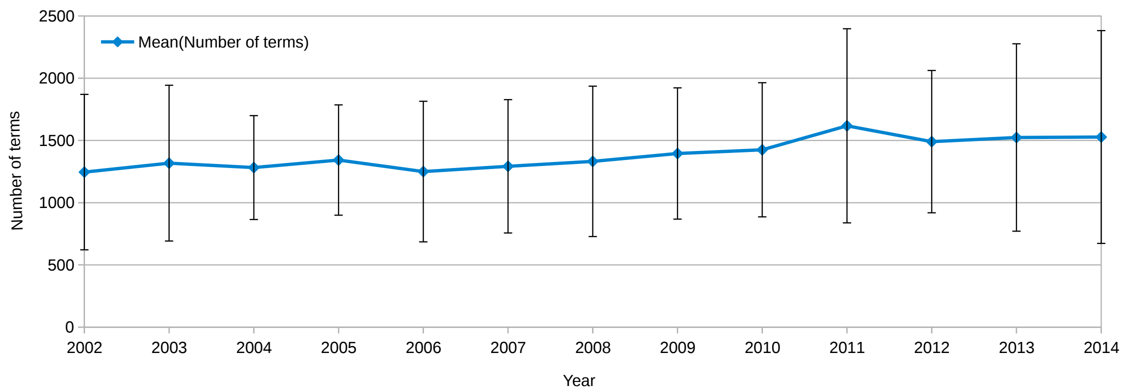

Figure 6.2: The average length of CEO letters over time. The error bars represent the standard deviation of the lengths.

6.1.1 Evolution of Sentiment in CEO Letters

CEO letters in annual reports typically reflect the previous year, the financial results and provide an outlook on the next year. Hence, they are not solely forward-looking, but very well suited for analyzing how the sentiment of managers has been changing over time.

As Figure 6.2 shows, the average length of these documents increased slightly over time. From an average number of 1250 terms in 2002, the length increased to 1530 terms on average

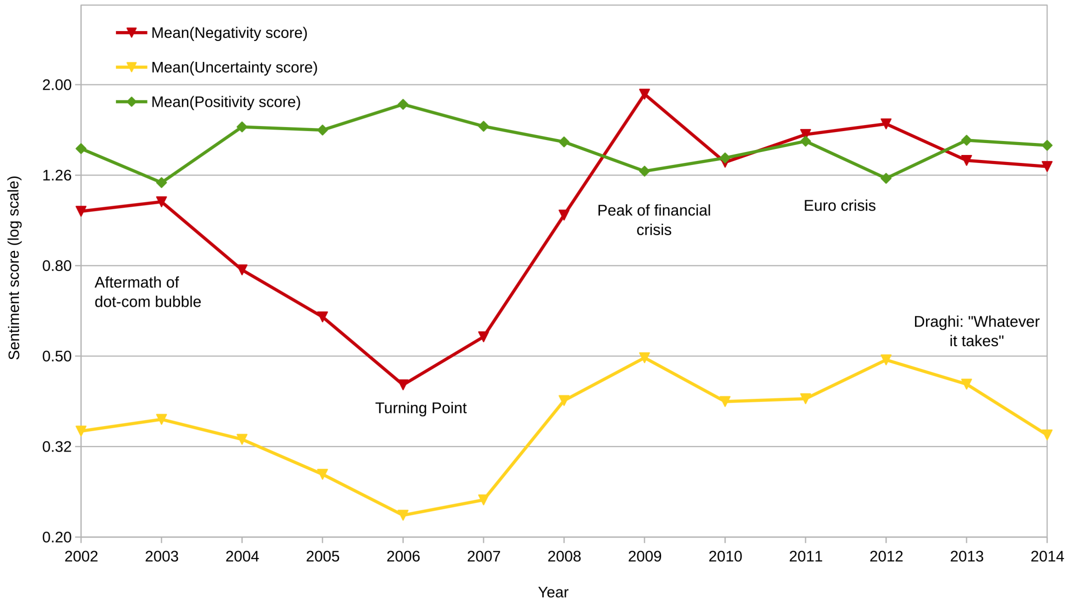

Figure 6.3: Evolution of positivity, negativity, and uncertainty in CEO letters over time.

in 2014[1]. The individual document lengths vary considerably: the shortest CEO letter consists of 329 terms[2], the longest one of 4134 terms[3]. This is also reflected by the *standard deviation* (SD)[4] of the lengths, which ranges from 417 to 855 terms.

Figure 6.3 shows how positivity, negativity, and uncertainty in CEO letters have been evolving since 2002. The individual data points represent the arithmetic mean of the document-level sentiment scores for each year. In 2002 and 2003, CEO letters contained more negative sentiment than in the following years. Banks might have emphasized that the recession following the burst of the *dot-com bubble* was still not over and that recovery had not yet arrived. Between 2003 and 2006, the letters became more positive and less negative from year to year. The turning point was in 2006—from that time on, negativity in CEO letters rose and quadrupled within three years. During the same period, positive sentiment scores decreased continuously. The summit of these evolutions was in 2009, in the midst of the financial crisis. The letters in 2010 had been already much more optimistic, but negativity increased in 2011 and 2012 again when CEOs recognized that the crisis was still not over.

The evolution of the uncertainty scores is similar to the negative sentiment scores. This observation is supported by a high correlation coefficient of 0.93, which is an almost perfect

[1]The number of terms measures the length of the original documents, i.e. before numbers, stop words and the like were removed.

[2]*Credit Mutuell*, published in 2008.

[3]*Piraeus Bank*, published in 2014.

[4]Standard deviation is defined as the positive square root of the variance (Wooldridge, 2002, p. 679). If the SD is low, the individual data points are relatively close to the mean, and vice versa.

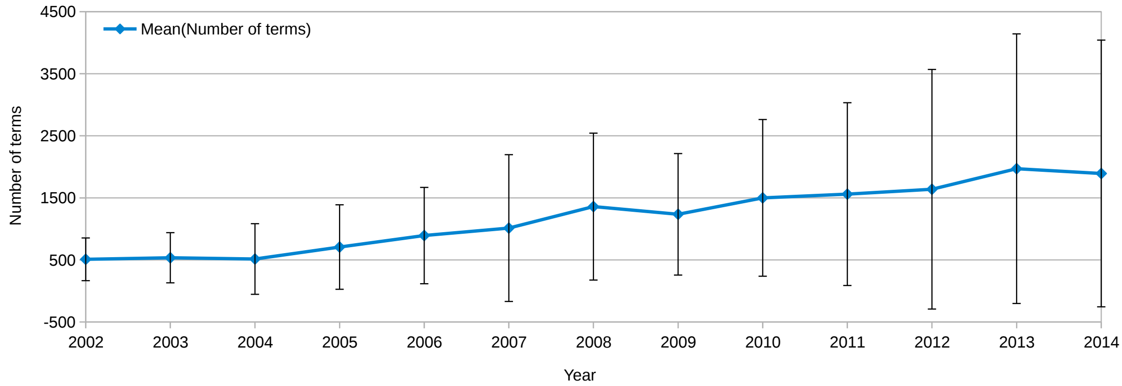

Figure 6.4: The average length of outlook sections over time. The error bars represent the standard deviation of the lengths.

positive linear correlation. Like the negativity score, uncertainty was still high in 2002 and 2003. The lowest average uncertainty score dates back to 2006, but the number more than doubled within the following three years. In 2009, uncertainty reached a first culmination point. The second peak could be observed in 2012, when economic recovery was still sluggish and doubts about the future of the common currency Euro came up. Since 2012, uncertainty has been decreasing quite sharply. Most likely, this can be attributed to an important and often-cited speech by ECB president Mario Draghi, who calmed the financial markets with the announcement to do "whatever it takes to preserve the euro. And believe me, it will be enough"[5].

Another observation is that the average uncertainty scores are much lower than the average positivity and negativity scores. A plausible interpretation thereof is that CEOs rather use clear statements than uncertain language.

6.1.2 Evolution of Sentiment in Outlook Sections

The average length of outlook sections between 2002 and 2014 is depicted in Figure 6.4. A first interesting observation is that both the number and the average length of outlook sections in annual reports have been increasing continuously since 2002[6]. A possible explanation for this fact is that banks became more concerned about the future and discuss different scenarios. Some banks also extended the section from a general to a more fine-grained outlook covering the prospects of different departments or business units. A second observation is that the standard deviation of the lengths also grew over time. The largest SD is 2172 terms and dates to 2013. A glance at the individual figures reveals that the lengths of outlook sections range from a short paragraph with 48 terms[7] to a document of 7 pages with 7628 terms[8].

[5]A transcript of this speech is available at `https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html`, accessed January 7th, 2015.

[6]The numbers of available outlooks sections are given in Table 5.2.1.

[7]*Pohjola Group*, published in 2005.

[8]*UniCredit Bank Austria*, published in 2013.

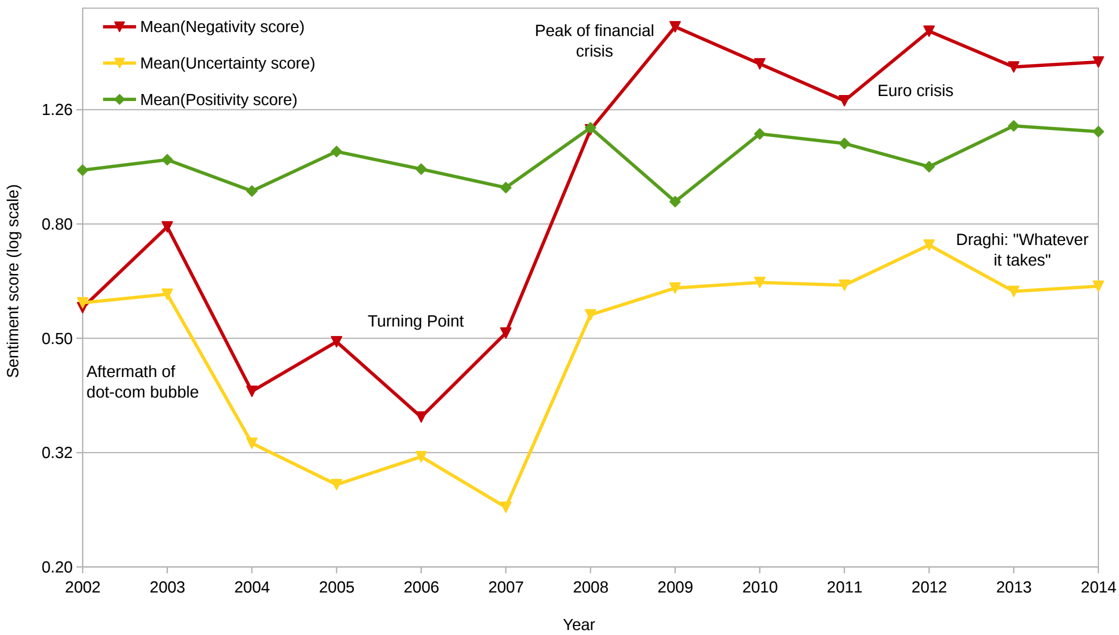

Figure 6.5: Evolution of positivity, negativity, and uncertainty in outlook sections over time.

As Figure 6.5 illustrates, negativity of the outlooks starts off at a relatively low level and had even lower values between 2004 and 2007. From 2006 on, negative sentiment kept increasing during the next years. Within just two years, negativity scores had more than tripled. This reflects considerable concerns of the banks about the situation in the financial markets[9]. Positivity was quite stable during all these years. Before 2008, it had much higher scores than negativity. After 2008, it was the opposite case—negativity scored higher than positivity.

Uncertainty was relatively high in 2002 and 2003, i.e. in the aftermath of the dot-com bubble. Between 2004 and 2007, banks had been very confident about the future. Uncertainty scores remained constant at a low level. This changed drastically in 2008, when the average score doubled within one year. It considerably increased again between 2011 and 2012. The high score in 2012 can primarily be attributed to the sovereign debt crisis in European countries and the lack of confidence about future economic growth. Since then, uncertainty slightly decreased, but the score in 2014 was still as high as in the aftermath of the dot-com bubble respectively during the financial crisis.

6.1.3 A Comparison of Two Groups of Countries in the Eurozone

The European countries Greece, Spain, Portugal, Italy, and Ireland, often referred to as *peripherals*[10], are struggling with economic challenges. They face weak economies, high unemployment,

[9] In spring 2008, the banks Bear Stearns and Northern Rock failed; see Section 2.1.1.

[10] Media and newspapers often use the acronym PIIGS for this group of countries, but it is not used here because of the derogatory character.

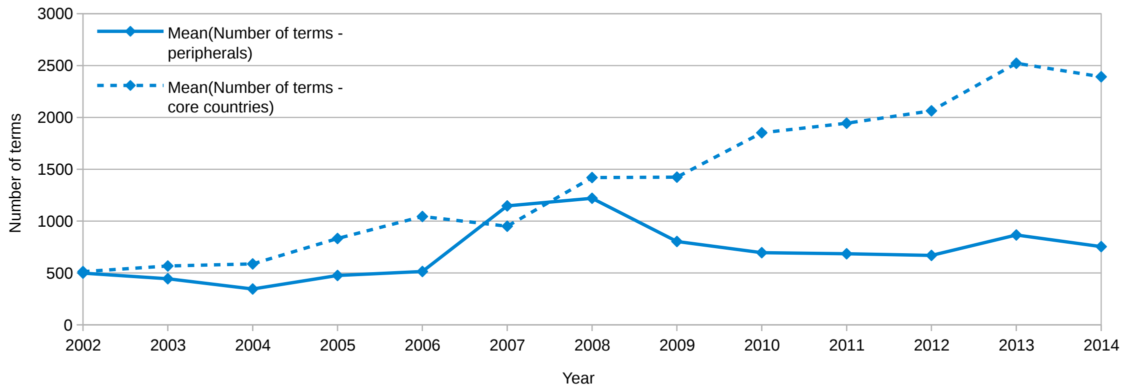

Figure 6.6: The average length of outlook sections in core and peripheral Eurozone countries.

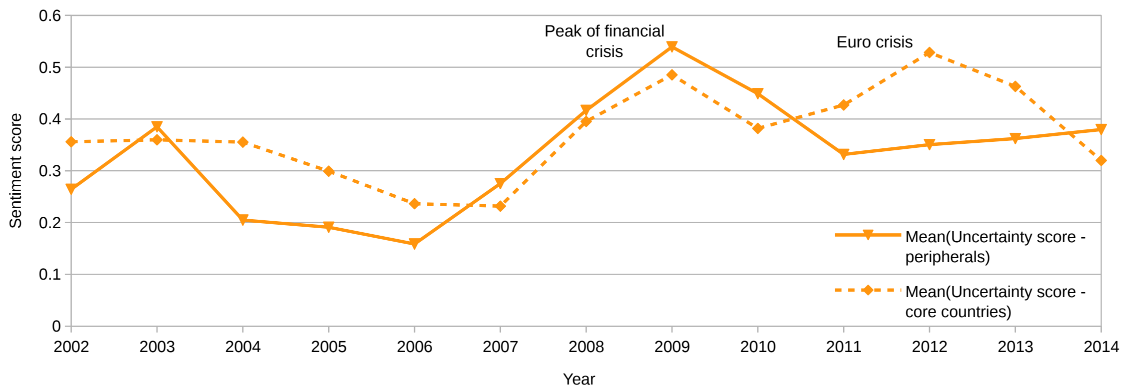

Figure 6.7: Evolution of uncertainty in CEO letters of core and peripheral Eurozone countries.

high government debts, and problems in their banking sectors. How do the collected sentiment scores reflect this situation?

A first observation is that the length of CEO letters remains fairly constant in both groups of countries, but CEOs in peripheral countries write about 30% longer texts. Regarding the outlook sections, Figure 6.6 reveals that the increased length of outlook sections (see also Figure 6.4) can mainly be attributed to banks in core countries.

The average levels of uncertainty in the CEO letters evolved in a similar fashion in both groups, see Figure 6.7. Until 2006, uncertainty in core countries was about equal or higher than in the peripherals. Between 2007 and the peak of the financial crisis in 2009, the scores increased in a parallel way. It is further interesting to observe that uncertainty in CEO letters of the peripherals does not reflect the Euro crisis around 2012, while the corresponding score of the core countries reached its highest value in that year. Hence, CEOs in core countries were more concerned about the future of the common currency Euro than those in the peripheral countries.

Figure 6.8 reveals that negativity scores based on the CEO letters of banks in peripheral

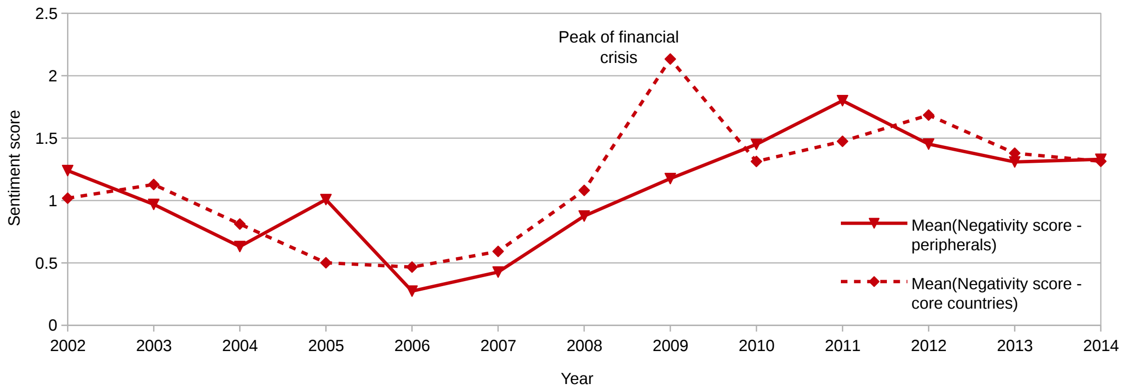

Figure 6.8: Evolution of negativity in CEO letters of core and peripheral Eurozone countries.

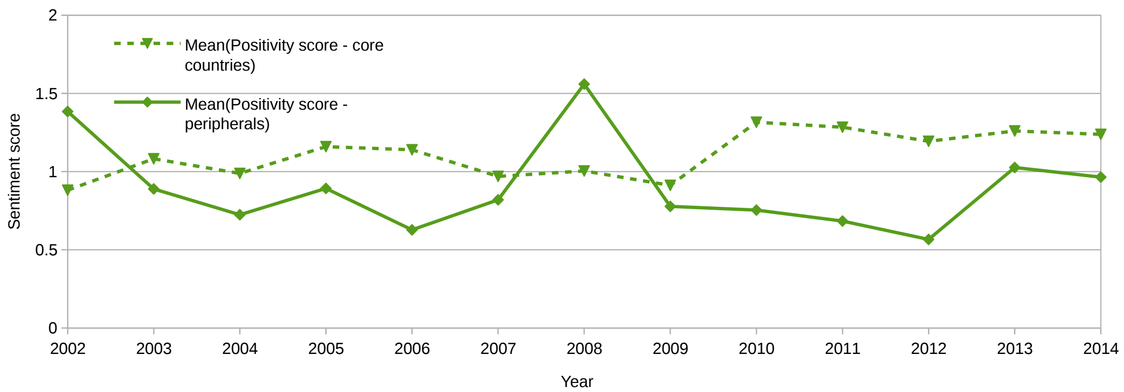

Figure 6.9: Evolution of positivity in outlook sections of core and peripheral Eurozone countries.

countries kept increasing between 2006 and 2011, but the peak of the financial crisis was not reflected in these scores. In contrast, scores for the core countries reached their maximum in 2009.

Another interesting fact can be observed in Figure 6.9, which depicts positivity scores in outlook sections. Positive sentiment in the peripheral countries reached a remarkable peak in 2008 and decreased rather quickly in the subsequent years. A glance at the individual figures shows that this peak cannot be attributed to outliers. Banks in the peripheral countries rather seemed to be overconfident about the future, but this changed when the crisis struck. In 2013, however, positivity scores recovered again.

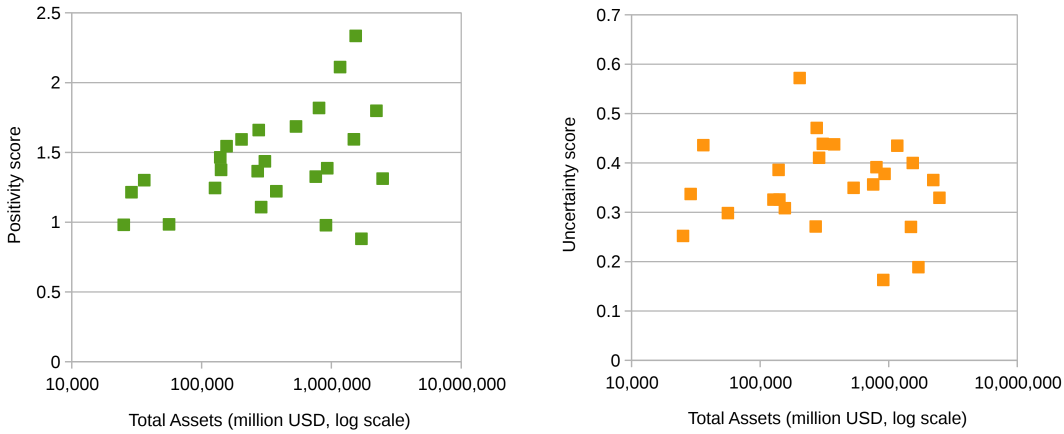

Figure 6.10: Bank size (measured by total assets) in comparison to uncertainty and positivity of CEO letters.

6.1.4 Is There a Correlation Between Bank Size and Language Used by CEOs?

The quantitative data set also contains the banks' *total assets* by the end of 2013. These figures serve as proxy for the size of banks. In this context, it is interesting to analyze whether CEOs of large banks use a different language than those of smaller ones. One could assume that managers of large banks are overconfident, i.e. they use a comparatively positive language, and do not focus on uncertainty because they are *too big to fail*[11].

Correlation coefficient	Uncertainty	Negativity	Positivity
Total Assets	-0.07	-0.01	0.57

Table 6.1: Correlation coefficients between total assets, uncertainty, negativity, and positivity.

The corresponding data points are depicted as scatter plots in Figure 6.10. The data in Table 6.1 show almost no correlation between bank size and uncertainty or negativity. Interestingly, there is a considerable correlation of $r = 0.57$ between bank size and positive sentiment. A statistical interpretation of this number is that 32 % of the positivity score's *variation* can be explained with the bank size ($r^2 = 0.32$). However, this does not confirm that the correlation is also statistically significant. Under the assumption of normally distributed variables, this can be examined via a *t-test*. The t-statistic (cf. Steiger, 1980) is calculated by

$$t = r\sqrt{\frac{n-2}{1-r^2}}. \tag{6.1}$$

The correlation coefficient is represented by r, and n is the sample's size. Positivity scores of CEO letters are available for $n = 25$ banks, so the t-statistic is $t = 3.32$ and the corresponding p-value is $p = 0.04$. Hence, the correlation between bank size and positivity is indeed statistically

[11] See Shull (2010) for a discussion of the term and Tracy (2014) for an example of a probably overconfident CEO.

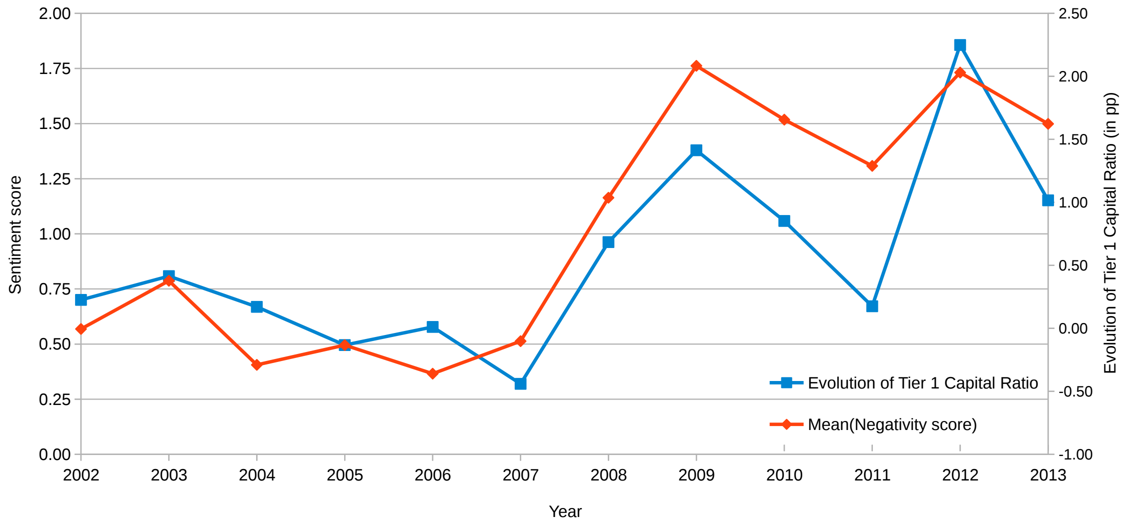

Figure 6.11: Evolution of the Tier 1 capital ratio compared to negativity in outlook sections.

significant if a 95 % confidence level is assumed. The null hypothesis of no correlation can be rejected. A limitation of this statement is that the figures were calculated on the bank size figures by the end of 2013. It might be more appropriate to use an average value from the years between 2002 and 2013.

6.1.5 Do Sentiment Scores Predict Quantitative Risk Measures?

The class labels for supervised classification are derived from the Tier 1 capital ratio. Comparing sentiment scores to the T1's average evolution in the same year reveals interesting relations. One of them is depicted in Figure 6.11: it shows both negativity scores in outlook sections and the T1 evolution since 2002. Note that sentiment scores are derived from outlook sections published between March and May of a year, and the corresponding T1 data points show the ratio's evolution until the end of that year.

The graphs in Figure 6.11 indicate a correlation between both data series. Tables 6.2 (for CEO letters) and 6.3 (for outlook sections) show that there are indeed strong correlations.

Correlation coefficient	Uncertainty	Negativity	Positivity
T1 evolution	0.86	0.79	-0.69

Table 6.2: Correlation coefficients between T1 evolution and sentiment scores of *CEO letters*.

It is interesting to analyze the data by a regression model for predicting the T1 evolution. A first question here is: which sentiment scores should be included in the model? According to the correlation coefficients, it is more meaningful to use the scores derived from outlook sections. Uncertainty and negativity in outlook sections are highly correlated ($r = 0.83$), so

Correlation coefficient	Uncertainty	Negativity	Positivity
T1 evolution	0.85	0.89	0.12

Table 6.3: Correlation coefficients between T1 evolution and sentiment scores of *outlook sections*.

Variable	Coeff.	Std. Err.	t-value	P>|t|
Mean(Positivity score)	-1,0096	0,8679	-1,1632	0,2747
Mean(Negativity score)	1,0597	0,1708	6,2031	0,0002
Intercept	0,4799	0,8642	0,5553	0,5922

Multiple R-Squared: 0,8134
Adjusted R-Squared: 0,7719

Figure 6.12: Results of a KNIME regression model based on positivity and negativity in outlook sections.

they should not be used together in a regression model because of the *multicollinearity* issue[12]. Hence, negativity and positivity are chosen as independent variables which try to explain the Tier 1 capital ratio's evolution.

Figure 6.12 depicts the regression model built by the KNIME *Linear Regression Learner* node. The coefficients can be interpreted as follows: an increase of the average positivity score in outlook sections by one unit results in a 1.0096 pp lower T1 evolution. On the other hand, if the negativity score rises by one unit, the T1 evolution increases by 1.0597 pp. If both negativity and positivity are zero, the Tier 1 capital ratio would increase by the computed intercept, which is 0.4799 pp.

The third column in Figure 6.12 lists the standard errors of the coefficients which are needed for constructing the test statistics. The last columns provide the t-statistic and the corresponding p-value. The latter is used for deciding whether the null hypothesis of the coefficient being zero holds. If a 95 % confidence level is assumed, the null hypothesis can be clearly rejected for the mean negativity score, but not for positivity.

It is therefore safe to assume that the positivity score is not meaningful in this model, so it can be rejected. Figure 6.13 shows the model with *negativity* as the only explaining variable. All coefficients are now statistically significant and can be explained by the following example.

Suppose that the T1 was 10 % at the end of a specific year. In March of the subsequent year, the bank publishes an outlook where the negativity score is zero[13]. The model would then predict a Tier 1 capital ratio of 9.498 % for the end of that year, because the T1 decreases by 0.502 pp. If, however, the negativity score is 1.3, the predicted T1 evolution would be $-0.502 + 0.9963 \times 1.3 = 0.79$. Hence, the T1 would increase to 10.79 %.

[12]Multicollinearity occurs when two or more independent variables in the model are highly correlated. Hence, the variables are not independent, but are rather able to explain each other. This decreases the regression model's quality (cf. Wooldridge, 2002, p. 95).

[13]The score would be zero if the outlooks do not contain words related to negativity, which is unlikely.

Variable	Coeff.	Std. Err.	t-value	P>|t|
Mean(Negativity score)	0,9963	0,1647	6,0478	0,0001
Intercept	-0,502	0,1883	-2,6651	0,0237

Multiple R-Squared: 0,7853
Adjusted R-Squared: 0,7638

Figure 6.13: Regression model based on negativity in outlook sections.

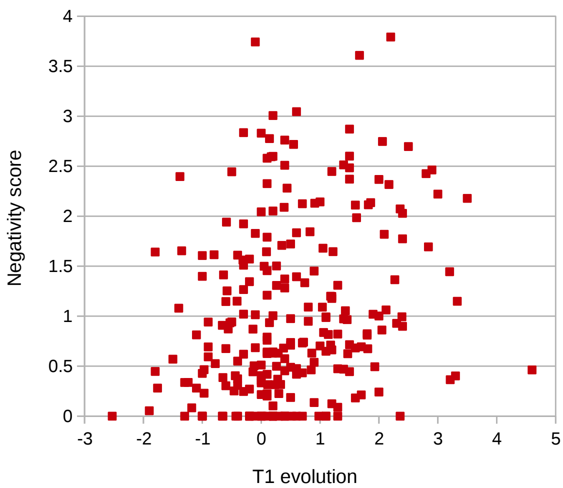

Figure 6.14: Individual negativity scores in outlook sections compared to the Tier 1 capital ratio's evolution.

Based on the outlooks published in spring 2014, the model predicts an increase of the average Tier 1 capital ratio by 0.96 pp until the end of 2014. This figure cannot be validated in this thesis because the data had not been published by February 2015. However, the predicted and actual T1 evolutions can be compared if one trains the model without the T1 data for 2013. In this case, the model predicts an increase of the T1 of 1.004 pp for 2013, while the actual value was 0.92 pp.

According to the *adjusted R-squared*, about 76 % of the T1 evolution's variation can be explained by the negativity score. In contrast to the *multiple R-squared*, this figure includes a penalty for additional independent variables in a model. This is done because the multiple R-squared cannot decrease if new variables are added—although the additional variables might not improve the model. More details on these measures are provided in Wooldridge (2002, p. 192ff).

A model of similar quality could be constructed by analyzing uncertainty in outlook sections. However, a drawback of all regression models is that they cannot model *external shocks* which influence the T1 evolution, but are not adequately covered in the outlook sections. Examples for such shocks would be new regulations concerning the minimum capital ratio or monetary policy actions by the ECB.

It should also be emphasized that these regression models are based on *aggregated* figures. Applying them on *individual* banks could lead to wrong conclusions. This assumption is supported by Figure 6.14, which compares individual negativity scores of individual outlook sections with the associated T1 evolutions. Although it is still possible to identify a positive relationship between the variables, the variance is too big for a meaningful regression model[14].

This phenomenon is known as *ecological correlation*[15]. Although it is theoretically possible that individual figures have the same correlations as the aggregates of these figures, this is very unlikely in practice. As William S. Robinson puts it: "From a practical standpoint, therefore, the only reasonable assumption is that an ecological correlation is almost certainly not equal to its corresponding individual correlation" (Robinson, 1950, p. 357).

6.2 Evaluation of the Supervised Classification Approach

The supervised classification experiment aims to assess whether it is possible to predict the T1 evolution based on CEO letters or outlook sections. The class labels *UP* and *DOWN* have been assigned according to the direction of the T1 evolution.

Overview on the Results. Table 6.4 gives an overview of the experiments and lists the respective performance measures. In contrast to the lexicon-based experiment, the main focus of the supervised classification approach lies on predicting the T1 trend for *individual* banks. An analysis of the data in the table reveals interesting facts. First, feature selection based on document frequency and information gain works better than the approach based on word lists. Second, the classifiers trained with CEO letters yield better results than the ones trained with outlook sections. Finally, three out of the four SVM results are not meaningful due to the following reason: the parameter optimization of C suggests to choose a very low value, which indeed maximizes the classifier accuracy—but these SVMs simply assign the class *UP* to every instance. However, the remaining SVM clearly yields the best results among the employed algorithms.

The non-meaningful classifiers can be seen as a baseline for comparisons since they deliver the same results as a simple *majority classifier*. The latter assigns all instances to the dominating class in the sample. Hence, classification is done by analyzing the class distribution in the training set and not by analyzing the text data. In the present data set, such a classifier achieves an accuracy of about 70 %, which equals the share of *UP*. A good supervised classifier needs to yield much better results.

6.2.1 Classification based on Topic-Specific Sentiment Words

At the aggregate level, the analysis based on topic-specific sentiment words delivered very promising results (see Section 6.1). However, as Figure 6.14 implies, this does not necessarily hold for the individual level. The performance measures in Table 6.4 confirm this assumption: none of the classifiers based on feature selection method (1) is able to outperform the baseline. Both SVMs simply classify every instance as *UP*, and the Naïve Bayes classifiers also deliver

[14] If the regression model is built with non-aggregated data, it explains only 6.6 % of the variation.

[15] It is also known as *ecological fallacy*.

Feature selection	Document type	Classifier	Accuracy	Precision U	Recall U	Precision D	Recall D
(1) based on topic-specific sentiment words	CEO letters	NB	0.703	0.741	0.889	0.500	0.263
		SVM	0.703	0.703	**1.000**	n.a.	0.000
	Outlook sections	NB	0.563	0.685	0.703	0.246	0.230
		SVM	0.703	0.703	**1.000**	n.a.	0.000
(2) based on document frequency and information gain	CEO letters	NB	0.750	0.774	0.911	0.636	0.368
		SVM	**0.792**	**0.810**	0.919	**0.718**	**0.491**
	Outlook sections	NB	0.704	0.704	**1.000**	n.a.	0.000
		SVM	0.704	0.704	**1.000**	n.a.	0.000

Table 6.4: Overview on the results of the supervised classification experiment. Bold numbers indicate the best results, U class *UP*, and D class *DOWN*.

unsatisfactory results. The NB classifier which analyzes outlook sections performs even below the baseline. Table 6.5 shows the confusion matrix for this NB classifier: 102 respectively 14 instances were correctly classified, but 47 instances were assigned to class *DOWN* although they actually belong to class *UP*. Furthermore, the NB classifier predicted class *UP* for 43 instances although their true class is *DOWN*. The conclusion for this part of the experiment is that sentiment words do not have any discriminatory power at the level of individual bank disclosures.

		True class	
		UP	DOWN
Predicted class	UP	102	43
	DOWN	47	14

Table 6.5: Confusion matrix for NB classification of outlook sections with feature selection approach (1).

6.2.2 Classification based on Document Frequency and Information Gain

The second part of the supervised classification experiment selects the relevant features by removing rare words and calculating the information gain of the remaining ones. This approach aims to reduce noise by using only features with considerable discriminatory power. It turns out that most words do not have any discriminatory power at all. Hence, their information gain is close to zero. Table 6.6 provides the short, but exhaustive list of stemmed words with an information gain greater than zero. It is not surprising that the list contains stems which are directly related to the financial crisis, e.g. *recess* and *crisi*.

As Table 6.4 shows, this approach for feature selection achieves better results than the first one, but only when the classifiers are trained with the CEO letter corpus. Most likely, this can be explained with the fact that outlook sections provide less terms with discriminatory power than CEO letters. Naïve Bayes correctly classifies 75 % of the instances, and the optimized SVM[16] yields 79.2 %. The other SVM performance measures can be interpreted as follows: 81 % of

[16]The optimum value for the parameter C turned out to be 111.

CEO letters		Outlook sections	
Word Stem	**IG**	**Word Stem**	**IG**
excel	0.077	impact	0.055
recess	0.072	crisi	0.052
crisi	0.067		
difficult	0.063		
affect	0.063		
condit	0.063		
month	0.060		

Table 6.6: The information gain of word stems in CEO letters and outlook sections.

the instances classified as *UP* were indeed instances where the Tier 1 capital ratio increased (= precision U). Furthermore, the SVM correctly identified almost 92 % of the instances which belong to the class *UP* (= recall U).

The figures for *UP* are satisfactory, but it might be more interesting for banking supervisors to identify banks where the T1 will be *decreasing* throughout the year. Unfortunately, the figures for *DOWN* are worse than those for *UP*. 71.8 % of the instances classified as *DOWN* indeed belong to this class (= precision D), but the SVM correctly identified only about half of the instances with decreasing T1 (= recall D). It would be possible to increase recall for the class *DOWN* by utilizing a *cost sensitive classifier*[17]. However, this would worsen the classifier's accuracy and precision.

		True class	
		UP	DOWN
Predicted class	UP	124	11
	DOWN	29	28

Table 6.7: Confusion matrix for SVM classification of CEO letters with feature selection approach (2).

Table 6.7 provides the confusion matrix for the best SVM classifier. It shows that 124 instances of class *UP* respectively 28 instances of class *DOWN* were correctly classified. Eleven instances were classified as *UP*, but in fact the Tier 1 capital ratio *decreased*. Finally, the SVM predicted for 29 instances a decreasing T1 although they actually belong to class *UP*.

These results are better than the baseline and demonstrate a noticeable potential for supervised classification even at the level of individual bank disclosures. Nevertheless, they are not good enough for reliable predictions. There are two reasons for this improvable performance: on the one hand, text mining respectively sentiment analysis has its limitations in terms of cap-

[17] A cost sensitive classifier penalizes specific errors. For example, it could be configured to decrease the share of false positives.

turing the actual meaning of a text. On the other hand, banks cannot completely control the Tier 1 capital ratio because of the dynamics in financial markets—and even if they could, annual reports would not contain all the relevant information.

6.2.3 Aggregation of Individual Classification Results

In the lexicon-based approach, the analysis based on aggregated data found promising results. An interesting question is what the results of the supervised classification reveal if they are aggregated by year.

Table 6.8 shows how the true and predicted classes are distributed in each year. The percentages were calculated by counting the respective classifications and dividing them by the total number of instances in that year. The *majority* columns indicate which class dominated in the corresponding year.

	True classes			**Predicted classes**		
Year	**UP (%)**	**DOWN (%)**	**Majority**	**UP (%)**	**DOWN (%)**	**Majority**
2002	66.7	33.3	UP	77.8	22.2	UP
2003	92.3	7.7	UP	84.6	15.4	UP
2004	66.7	33.3	UP	83.3	16.7	UP
2005	42.9	57.1	DOWN	50.0	50.0	n.a.
2006	40.0	60.0	DOWN	33.3	66.7	DOWN
2007	29.4	70.6	DOWN	41.2	58.8	DOWN
2008	81.8	18.2	UP	95.5	4.5	UP
2009	95.0	5.0	UP	95.0	5.0	UP
2010	88.2	11.8	UP	94.1	5.9	UP
2011	56.3	43.8	UP	75.0	25.0	UP
2012	100.0	0.0	UP	100.0	0.0	UP
2013	71.4	28.6	UP	90.5	9.5	UP
2014	n.a.	n.a.	n.a.	86.4	13.6	UP

Table 6.8: Overview on the distribution of true and predicted classes per year. The underlying data are based on SVM classification of CEO letters with feature selection approach (2).

Although the class distributions differ considerably in some years, there is always the same tendency. Hence, the aggregated classification data accurately predict whether the majority of banks will increase or decrease their Tier 1 capital ratio in the following year. An exception is the year 2005, where none of the predicted classes has a majority. For the unpublished figures of 2014, the data clearly predict an increase of the T1 in most banks. This finding is in line with the linear regression approach, which predicts that the average T1 increases by 0.96 pp in 2014 (see Section 6.1.5).

CHAPTER 7

Concluding Remarks

The last chapter provides a summary of this thesis and outlines its main findings in order to answer the research questions. Furthermore, it highlights opportunities for future research in the area of risk sentiment analysis.

7.1 Summary and Main Findings

The financial crisis of 2007-08 highlighted the need for reforms in the banking sector, including more effective financial supervision. In this context, the *European Central Bank* (ECB) started to implement the *Single Supervisory Mechanism* (SSM), which is a part of the European Banking Union. It aims to enhance safety and soundness of credit institutions and transfers supervisory tasks from the national authorities to the ECB.

The ECB is now responsible for the supervision of 123 large banks in the Eurozone. In order to fulfill its tasks adequately, the ECB should utilize a range of information sources. This includes not only quantitative data, but also unstructured information like textual disclosures. In this context, this thesis explored how banking supervisors could utilize *sentiment analysis* for risk assessments. While the primary purpose of this popular text mining technique is to classify documents like product reviews according to their polarity, this work utilizes it for measuring a bank's attitude and opinions about risk.

Suitable Data Sources. Among the tackled challenges were the following tasks: first, proper data sources had to be identified. In general, suitable documents have to be publicly accessible and need to contain forward-looking information about risk. Furthermore, an English version has to be available and it needs to be published at least annually.

The analysis of potential document types revealed that two sections in a bank's annual report are particularly well suited for this thesis, namely the *CEO letter* and the *outlook section*. Both fulfill the outlined criteria. CEO letters represent the *tone from the top* and provide subjective information about the bank's current and future situation. Outlook sections are exclusively forward-looking and reveal opinions about the near future.

As a supplement to text data, the analyses in this thesis required appropriate quantitative risk indicators which should represent the bank's financial health and its general risk exposure. In addition, they need to fulfill the same general criteria as the text data. It turned out that the *Tier 1 capital ratio* (T1) is the best suited quantitative risk indicator. The T1 sets the most secure forms of bank capital in relation to its risk-weighted assets and is widely used in banking supervision, e.g. as a key ratio for the ECB's stress test in fall 2014.

Realization of the Experiments. The collected data had to be processed and analyzed. First of all, it was necessary to extract the relevant sections from annual reports and to prepare the document corpus. The open source data mining tool *KNIME* was then used for conducting the experiments.

The first experiment was a *lexicon-based analysis* which derived sentiment scores for each document. In particular, the analysis started with tagging uncertain, negative, and positive words according to a topic-specific lexicon for the financial domain. Further steps included simple negation handling and a term weighting approach. The latter weighs terms according to their frequency in an individual document and their distribution in the corpus. In addition, term weighting accounts for different document lengths. Finally, the document scores were calculated by aggregating the individual sentiment scores.

The second experiment employed *supervised classification* in order to predict the future evolution of quantitative risk indicators. After reading and parsing the data, a linguistic preprocessing step removed irrelevant words and other elements from the documents. The subsequent node contained two different methods for *feature selection*: the first one assumes that the relevant features are defined via the lexicon from the previous experiment. The second approach incorporates the feature selection techniques *document frequency* and *information gain*. The second experiment's last step is to classify the selected features with two common machine learning algorithms, namely *Naïve Bayes* and *Support Vector Machine*.

Main Findings. The conducted experiments set the ground for various analyses. While both experiments revealed interesting results, the lexicon-based approach provided the best insights. Its aggregated sentiment scores were primarily used for analyzing how risk sentiment has been evolving since 2002, especially during the financial crisis. Furthermore, the figures were employed within regression models. Here is an overview on the main findings of the lexicon-based approach:

- Uncertainty and negativity evolved in a similar way over time. The graphs clearly reflect major events between 2002 and 2014, in particular the aftermath of the dot-com bubble, the confidence around 2006, the financial crisis, and the calming effects of Mario Draghi's "Whatever it takes" speech. These observations hold for both CEO letters and outlook sections.

- Although these scores evolved similarly, it is noteworthy that the level of uncertainty is much lower than the levels of negativity or positivity. This indicates that CEOs rather use clear statements than uncertain language.

- There is a positive correlation between bank size, measured by total assets, and positive language used by CEOs. Although the correlation is not very strong, the relationship is statistically significant.
- While the average length of CEO letters increased only slightly over time, the average outlook section more than tripled its length within the last decade. An interpretation is that banks became more concerned about the future. Interestingly, the increase in length can mainly be attributed to banks in the core Eurozone countries.
- There is a strong correlation between uncertainty, negativity, and the Tier 1 capital ratio's evolution over time. Hence, the sentiment scores could be used in regression models for predicting the T1 evolution. The best results are achieved by employing negativity in outlook sections as the explaining variable. Since this relationship only holds for figures aggregated by year, it cannot be used for individual banks. It should also be noted that this method is not meant to be used as a stand-alone estimator for the T1 evolution. Instead, it should be combined with other estimation methods.

In the supervised risk classification approach, machine learning algorithms were trained for predicting whether the Tier 1 capital ratio will increase or decrease within the next year. The corresponding class labels are *UP* and *DOWN*. This approach primarily utilizes data at the level of individual banks. The main findings are as follows:

- Feature selection based on document frequency and information gain works better than the one based on word lists.
- In comparison to outlook sections, CEO letters contain more terms with an *information gain* greater than zero, and classifiers trained with these documents yield a better performance. This is surprising since CEO letters are—in contrast to outlook sections—not exclusively forward-looking.
- The best results were achieved with a SVM classifier. It correctly classifies 79.2 % of the instances. This is not outstanding if one considers that it is possible to yield an accuracy of 70 % simply by assigning the class *UP* to every instance.
- If the results of the best SVM classifier are aggregated by year, it is possible to accurately predict whether the majority of banks will increase or decrease their Tier 1 capital ratio in the following year.

The described systems have the potential to provide valuable insights for banking supervisors. However, it has to be emphasized that predictions for individual banks are relatively inaccurate and therefore not reliable. A better option is to use these techniques for macroprudential analyses[1]. Examples are predictions for the average Tier 1 capital ratio's evolution in the whole Eurozone or in groups of countries. Another option is to improve existing risk prediction frameworks by integrating sentiment scores derived from the textual disclosures.

[1]The term *macroprudential* became popular in the course of the financial crisis. It refers to the promotion of stability in the whole financial system. A related term is *microprudential*, which refers to the stability of individual banks (cf. Clement, 2010; de Larosière et al., 2009).

7.2 Opportunities for Future Research

The analyses in this thesis revealed interesting opportunities for future research in the area of risk sentiment analysis. This involves methodological improvements, but also expansions of the approach:

- It would be interesting to validate the results by conducting the study on a larger scale. One could incorporate data from all European banks, or from other regions. Of course, such a study would require fully automated retrieval and processing of the data.
- This thesis employed well-proven text mining methods for the experiments. Further research could utilize more advanced techniques, e.g. for detecting opinions which are expressed in a subtle manner.
- Another refinement of the sentiment score calculation would be different weightings, e.g. when it is desired that the risk sentiment of a large bank should have a greater weight than the one of a small bank.
- Besides using risk sentiment analysis as a stand-alone tool, one could incorporate the derived data in existing frameworks for risk assessments. Although the results are too inaccurate at the level of individual banks, they could still improve such frameworks[2]. It would be interesting to test whether this hypothesis holds.
- It turned out that sentiment scores quite accurately reflect the economic situation and major events. By expanding the approach to other industries or all major companies in a country, a *CEO Sentiment Indicator* could be constructed for supplementing the EU's *Economic Sentiment Indicator* (ESI)[3].

[2] For example, *Google Flu Trends* as a stand-alone tool overestimates the number of flu diseases, but combining Google's data with official health data yields the best predictions (cf. Lazer et al., 2014).

[3] See e.g. `http://www.oenb.at/en/Statistics/Standardized-Tables/Economic-and-Industry-Indicators/Economic-Indicators/Economic-Sentiment-Indicator-for-the-Euro-Area.html`, accessed January 28th, 2015.

APPENDIX A

Appendix

A.1 Document Corpus

The parsed document corpus is available online via the following website: `http://ifs.tuwien.ac.at/~mucke/BankRiskSAdata/`. The compressed files are protected and can be extracted with the password *thesis_nopp*.

Table A.1 and Table A.2 indicate whether the CEO letter respectively the outlook section for a specific bank and year was available and therefore integrated in the analysis. The sign ✓ denotes the document's availability, ✗ the opposite.

Availability of CEO Letters		Year													
Full Bank Name	**Country**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Sum**
AIB Group	GB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Banca Monte dei Paschi di Siena SpA	IT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	**0**
Banco Bilbao Vizcaya Argentaria SA	ES	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	**8**
Banco Popolare - Società Cooperativa	IT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	**0**
Banco Santander SA	ES	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Banque et Caisse d’Epargne de l’Etat Luxembourg	LU	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	**1**
BNP Paribas	FR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Caixa Geral de Depositos	PT	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	✓	**6**
Commerzbank AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Rabobank Nederland	NL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Credit Mutuel	FR	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	**4**
Deutsche Bank AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Dexia	BE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
DZ Bank Group	DE	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
Erste Group Bank AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Hypo Alpe-Adria Bank International AG	AT	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	✓	✓	**7**
ING Groep NV	NL	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
KBC Bank NV	BE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Landesbank Baden-Württemberg	DE	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	**7**
Landesbank Berlin Holding AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	**12**
Oesterreichische Volksbanken AG	AT	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
OP-Pohjola Group	FI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Piraeus Bank SA	GR	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	**9**
Raiffeisen Zentralbank Oesterreich AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Société Générale	FR	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	**3**
UniCredit Bank Austria AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
UniCredit SpA	IT	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	**8**
Sum		**15**	**19**	**17**	**19**	**19**	**23**	**25**	**21**	**20**	**21**	**20**	**22**	**22**	**257**

Table A.1: An overview on the availability of CEO Letters. Note: ✓ indicates *available*, ✗ stands for *not available*.

Availability of Outlook Sections		Year													
Full Bank Name	**Country**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Sum**
AIB Group (UK)	GB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Banca Monte dei Paschi di Siena SpA	IT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Banco Bilbao Vizcaya Argentaria SA	ES	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
Banco Popolare - Società Cooperativa	IT	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
Banco Santander SA	ES	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	**8**
Banque et Caisse d'Epargne de l'Etat Luxembourg	LU	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	**9**
BNP Paribas	FR	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	**5**
Caixa Geral de Depositos	PT	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	**6**
Commerzbank AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Rabobank Nederland	NL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Credit Mutuel	FR	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	**8**
Deutsche Bank AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Dexia	BE	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	**0**
DZ Bank Group	DE	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	**11**
Erste Group Bank AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Hypo Alpe-Adria Bank International AG	AT	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	**8**
ING Groep NV	NL	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
KBC Bank NV	BE	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	**6**
Landesbank Baden-Württemberg	DE	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	**7**
Landesbank Berlin Holding AG	DE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Oesterreichische Volksbanken AG	AT	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
OP-Pohjola Group	FI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Piraeus Bank SA	GR	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	**9**
Raiffeisen Zentralbank Oesterreich AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
Société Générale	FR	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**12**
UniCredit Bank Austria AG	AT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**13**
UniCredit SpA	IT	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**11**
Sum		**14**	**19**	**20**	**20**	**21**	**22**	**23**	**23**	**23**	**23**	**23**	**24**	**23**	**278**

Table A.2: An overview on the availability of outlook sections. Note: ✓ indicates *available*, ✗ stands for *not available*.

A.2 Quantitative Bank Data

Table A.3 lists the Tier 1 capital ratios for a specific bank at the end of a specific year. In Table A.4, further data like the banks' total assets and their world ranks are provided.

Tier 1 Capital Ratio (in %)		Year													
Bank Name	**Country**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**Average**
AIB Group	GB	6.50	6.90	7.10	8.20	7.20	7.09	6.53	7.39	8.30	9.40	14.60	16.90	18.50	**9.59**
Banca Monte dei Paschi di Siena SpA	IT	5.77	6.10	6.50	6.70	6.51	6.53	6.10	5.13	7.52	8.37	10.30	9.52	10.62	**7.36**
Banco Bilbao Vizcaya Argentaria SA	ES	8.50	8.40	8.50	7.90	7.50	7.80	6.80	7.90	9.36	10.54	9.90	10.25	11.72	**8.85**
Banco Popolare - Società Cooperativa	IT	6.59	7.30	7.90	7.60	7.40	7.70	5.17	6.39	7.69	7.16	8.34	11.18	10.60	**7.77**
Banco Santander SA	ES	8.44	8.00	8.26	7.16	7.88	7.42	7.71	8.75	9.97	9.86	10.91	11.17	12.60	**9.09**
Banque et Caisse d'Epargne de l'Etat Luxembourg	LU	n.a.	n.a.	9.80	10.90	9.30	9.30	9.40	14.60	14.40	13.10	14.40	15.20	16.60	**12.45**
BNP Paribas	FR	7.30	8.10	9.40	7.50	7.60	7.40	7.30	7.80	10.10	11.40	11.60	13.60	12.80	**9.38**
Caixa Geral de Depositos	PT	n.a.	n.a.	n.a.	n.a.	7.40	7.40	6.70	7.00	8.50	8.89	9.03	11.20	11.30	**8.60**
Commerzbank AG	DE	6.00	7.30	7.30	7.50	8.00	6.70	6.90	10.10	10.50	11.90	11.10	13.10	13.50	**9.22**
Rabobank Nederland	NL	10.20	10.30	10.80	10.90	11.60	10.70	10.70	12.75	13.80	15.70	17.00	17.20	16.60	**12.94**
Credit Mutuel	FR	n.a.	8.80	9.60	10.50	10.20	9.96	9.30	9.80	11.00	11.50	11.20	14.50	14.50	**10.91**
Deutsche Bank AG	DE	8.10	9.60	10.00	8.60	8.70	8.50	8.60	10.10	12.60	12.30	12.90	15.10	16.90	**10.92**
Dexia	BE	9.30	9.30	9.90	10.00	10.30	9.80	9.10	10.60	12.30	13.10	7.60	19.90	21.40	**11.74**
DZ Bank Group	DE	n.a.	5.80	7.00	7.90	8.10	9.70	7.70	7.40	9.90	10.60	10.10	13.60	16.40	**9.52**
Erste Group Bank AG	AT	6.20	6.30	6.30	6.70	6.80	6.60	7.00	7.20	9.20	10.20	10.40	11.60	11.80	**8.18**
Hypo Alpe-Adria Bank International AG	AT	n.a.	n.a.	n.a.	n.a.	n.a.	5.40	6.30	8.30	7.20	7.30	6.90	8.60	9.80	**7.48**
ING Groep NV	NL	7.00	7.31	7.59	7.70	7.30	7.60	5.80	7.30	7.80	9.60	9.60	12.00	11.70	**8.33**
KBC Bank NV	BE	8.80	8.80	9.50	10.10	9.45	8.45	7.80	9.60	10.92	12.40	11.60	13.80	16.20	**10.57**
Landesbank Berlin Holding AG	DE	5.70	5.60	6.10	7.50	8.10	7.20	6.70	7.76	8.50	10.12	9.81	12.17	11.87	**8.24**
Oesterreichische Volksbanken AG	AT	9.80	9.50	8.90	10.04	7.33	7.71	7.05	7.14	9.20	9.37	8.78	10.87	14.08	**9.21**
OP-Pohjola Group	FI	13.00	13.60	14.00	14.10	13.10	12.70	12.60	12.60	12.60	12.80	14.00	14.10	14.30	**13.35**
Piraeus Bank SA	GR	n.a.	n.a.	n.a.	10.30	8.80	7.40	9.76	8.00	9.10	8.70	n.a.	9.30	13.90	**9.47**
Raiffeisen Zentralbank Oesterreich AG	AT	7.20	7.40	7.50	9.20	8.30	9.00	8.80	7.00	9.40	9.30	9.90	11.40	10.40	**8.83**
Société Générale	FR	8.40	8.14	8.66	8.54	7.57	7.80	6.62	8.43	10.70	10.56	10.70	12.50	13.40	**9.39**
UniCredit Bank Austria AG	AT	7.80	6.80	7.80	7.85	8.29	11.62	8.20	6.82	8.68	10.35	10.90	10.80	11.60	**9.04**
UniCredit SpA	IT	n.a.	7.21	6.96	7.94	6.70	6.96	6.55	6.81	8.63	9.46	9.32	11.44	10.09	**8.17**
Landesbank Baden-Württemberg	DE	n.a.	n.a.	n.a.	n.a.	n.a.	7.40	6.50	6.90	9.80	11.40	12.90	15.30	18.30	**11.06**
Average		**7.93**	**8.03**	**8.49**	**8.81**	**8.38**	**8.22**	**7.69**	**8.50**	**9.91**	**10.57**	**10.92**	**12.83**	**13.76**	

Table A.3: An overview on the individual Tier 1 capital ratios and its average values for each year and bank. These data are the basis for the assigned class labels respectively the calculations regarding the T1 evolution and were retrieved from `https://bankscope.bvdinfo.com` on December 14th, 2014.

Bank Name	City	Country	Total Assets (million USD, end of 2013)	World Rank by Assets
AIB Group	Belfast	GB	25,156.457	735
Banca Monte dei Paschi di Siena SpA	Siena	IT	274,590.951	113
Banco Bilbao Vizcaya Argentaria SA	Bilbao	ES	803,440.899	44
Banco Popolare - Società Cooperativa	Verona	IT	173,828.022	155
Banco Santander SA	Madrid	ES	1,538,598.803	17
Banque et Caisse d'Epargne de l'Etat Luxembourg	Luxembourg	LU	56,149.634	381
BNP Paribas	Paris	FR	2,482,607.898	4
Caixa Geral de Depositos	Lisbon	PT	155,789.269	164
Commerzbank AG	Frankfurt am Main	DE	758,048.540	47
Rabobank Nederland	Utrecht	NL	929,718.653	34
Credit Mutuel	Paris	FR	908,313.330	36
Deutsche Bank AG	Frankfurt am Main	DE	2,222,314.148	8
Dexia	Brussels	BE	307,455.521	n.a.
DZ Bank Group	Frankfurt am Main	DE	533,689.142	68
Erste Group Bank AG	Vienna	AT	275,653.149	n.a.
Hypo Alpe-Adria Bank International AG	Klagenfurt	AT	36,158.599	n.a.
ING Groep NV	Amsterdam	NL	1,490,310.291	n.a.
KBC Bank NV	Brussels	BE	287,833.400	109
Landesbank Berlin Holding AG	Berlin	DE	141,272.927	n.a.
Oesterreichische Volksbanken AG	Vienna	AT	28,829.265	661
OP-Pohjola Group	Helsinki	FI	139,264.928	173
Piraeus Bank SA	Athens	GR	126,892.290	184
Raiffeisen Zentralbank Oesterreich AG	Vienna	AT	203,177.629	n.a.
Société Générale	Paris	FR	1,703,574.667	15
UniCredit Bank Austria AG	Vienna	AT	270,597.157	114
UniCredit SpA	Milano	IT	1,166,512.748	27
Landesbank Baden-Württemberg	Stuttgart	DE	376,585.600	n.a.

Table A.4: General data about the analyzed banks. The figures were retrieved from `https://bankscope.bvdinfo.com` on December 14th, 2014.

Bibliography

Aggarwal, C. C. and Zhai, C. X. (2012a). A Survey of Text Classification Algorithms. In Aggarwal, C. C. and Zhai, C. X., editors, *Mining Text Data*, pages 163–222. Springer Science+Business Media.

Aggarwal, C. C. and Zhai, C. X. (2012b). A Survey of Text Clustering Algorithms. In Aggarwal, C. C. and Zhai, C. X., editors, *Mining Text Data*, pages 77–128. Springer Science+Business Media.

Aggarwal, C. C. and Zhai, C. X. (2012c). An Introduction to Text Mining. In Aggarwal, C. C. and Zhai, C. X., editors, *Mining Text Data*, pages 1–10. Springer Science+Business Media.

Aiginger, K. (2009). The Current Economic Crisis: Causes, Cures and Consequences. WIFO Working Papers 341, Austrian Institute of Economic Research (WIFO). Retrieved April 14th, 2014, from `http://karl.aiginger.wifo.ac.at/fileadmin/files_aiginger/publications/2009/WP_2009_341_.PDF`.

Alpaydin, E. (2014). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 3rd edition.

Amernic, J., Craig, R., and Tourish, D. (2010). *Measuring and Assessing Tone at the Top Using Annual Report CEO Letters*. Institute of Chartered Accountants of Scotland, Edinburgh. Retrieved March 28th, 2014, from `http://eprints.port.ac.uk/12648/1/CRAIG_2010_pub_Bk_Measuring_and_assessing_tone_at_the_top_using_annual_report_CEO_letters.pdf`.

Armah, N. A. (2013). Big Data Analysis: The Next Frontier. *Bank of Canada Review for Summer 2013*, pages 32–39.

Austrian Financial Market Authority (2014). Annual Report of the Financial Market Authority 2013. Retrieved May 24th, 2014, from `http://www.fma.gv.at/fileadmin/media_data/1_Ueber_die_FMA/2_Publikationen/FMA_JB13_englisch_screen_final.pdf`.

Bank Austria Creditanstalt (2006). Annual Report 2005. Retrieved December 17th, 2014, from `http://www.bankaustria.at/files/e_ba-ca_gb05_internet_060407.pdf`.

Bank for International Settlements (2009). Revisions to the Basel II market risk framework. Retrieved February 12th, 2015, from http://www.bis.org/publ/bcbs148.pdf.

BAWAG P.S.K. (2014). Consolidated Annual Report 2013. Retrieved January 29th, 2015, from https://www.bawagpsk.com/linkableblob/BAWAGPSK/181360/3/consolidated-annual-report-2013-data.pdf.

Bessis, J. (2002). *Risk Management in Banking*. John Wiley & Sons Ltd, Chichester, United Kingdom, 2nd edition.

Blanchard, O. and Illing, G. (2009). *Makroökonomie*. Pearson Studium, 5th edition.

Boudt, K. and Thewissen, J. (2014). All Words Are Not Equal: Sentiment Dynamics and Information Content Within CEO Letters. Working Paper. Retrieved March 26th, 2014, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2326400.

Breuss, F. (2013). European Banking Union. WIFO Working Papers 454, Austrian Institute of Economic Research (WIFO). Retrieved May 16th, 2014, from http://fritz.breuss.wifo.ac.at/Breuss_European_Banking_Union_WIFO_WP_454_Sept_2013.pdf.

Brunnermeier, M. K. (2009). Deciphering the Liquidity and Credit Crunch 2007-2008. *Journal of Economic Perspectives*, 23(1):77–100.

Cabanillas, L. G. and Ruscher, E. (2008). The Great Moderation in the euro area: What role have macroeconomic policies played? European Economy - Economic Papers 331, Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission. Retrieved April 26th, 2014, from http://ec.europa.eu/economy_finance/publications/publication12753_en.pdf.

Cannata, F., Casellina, S., and Guidi, G. (2012). *Inside the labyrinth of Basel risk-weighted assets: how not to get lost*. Number 132 in Questioni di Economia e Finanza. Banca d'Italia. Retrieved September 20th, 2014, from http://www.bancaditalia.it/pubblicazioni/qef/2012-0132/QEF_132.pdf.

Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems*. PhD thesis, Yale University.

Casu, B. and Chiaramonte, L. (2011). Are CDS spreads a good proxy of bank risk? Evidence from the financial crisis. *BANCARIA*, 11:82–98.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159.

Chen, J. (2014). Measuring Market Risk Under Basel II, 2.5, and III: VaR, Stressed VaR, and Expected Shortfall. *AESTIMATIO, The IEB International Journal of Finance*, 8:184–201.

Chen, X., Ye, Y., Williams, G., and Xu, X. (2007). A Survey of Open Source Data Mining Systems. In Washio, T., Zhou, Z.-H., Huang, J., Hu, X., Li, J., Xie, C., He, J., Zou, D., Li, K.-C., and Freire, M., editors, *Emerging Technologies in Knowledge Discovery and Data Mining*, volume 4819 of *Lecture Notes in Computer Science*, pages 3–14. Springer Berlin Heidelberg.

Chisholm, E. and Kolda, T. G. (1999). New Term Weighting Formulas For The Vector Space Method In Information Retrieval. Technical Report TN 37831-6367, Oak Ridge National Laboratory. Retrieved November 6th, 2014, from http://www.sandia.gov/~tgkolda/pubs/pubfiles/ornl-tm-13756.pdf.

Clement, P. (2010). The Term "Macroprudential": Origins and Evolution. *BIS Quarterly Review for March 2010*, pages 59–67. Retrieved January 28th, 2015, from http://www.bis.org/publ/qtrpdf/r_qt1003h.pdf.

Collignon, S. (2012). Europe's Debt Crisis, Coordination Failure, and International Effects. ADBI Working Papers 370, Asian Development Bank Institute. Retrieved May 8th, 2014, from http://www.adbi.org/files/2012.07.11.wp370.europe.debt.crisis.effects.pdf.

Commerzbank AG (2008). Annual Report 2008. Retrieved November 25th, 2014, from https://www.commerzbank.de/media/en/aktionaere/service/archive/konzern/2009_2/coba_gb2008_2.pdf.

Council of the EU (2013). Council Regulation (EU) 1024/2013 of 15 October 2013 conferring specific tasks on the European Central Bank concerning policies relating to the prudential supervision of credit institutions. *Official Journal of the European Union*, L287:63–89.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

de Larosière, J., Balcerowitz, L., Issing, O., Masera, R., Mc Carthy, C., Nyberg, L., Pérez, J., and Ruding, O. (2009). Report of the High-Level Group on Financial Supervision in the EU. Retrieved May 16th, 2014, from http://ec.europa.eu/internal_market/finances/docs/de_larosiere_report_en.pdf.

Deloitte LLP (2014). The Single Supervisory Mechanism (SSM): The big data issue. Retrieved January 29th, 2015, from http://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-fs-the-big-data-issue.pdf.

European Banking Authority (2013). EBA Final Draft Implementing Technical Standards. Retrieved April 15th, 2014, from http://www.eba.europa.eu/documents/10180/359626/EBA+ITS+2013+02+%28Draft+ITS+on+supervisory+reporting%29.pdf.

European Banking Authority (2014a). Results of 2014 EU-wide stress test. Retrieved November 17th, 2014, from https://www.eba.europa.eu/documents/10180/669262/2014+EU-wide+ST-aggregate+results.pdf.

European Banking Authority (2014b). Risk Assessment of the European Banking System - June 2014. Retrieved July 27th, 2014, from https://www.eba.europa.eu/documents/10180/556730/EBA+Risk+Assessment+Report+June+2014.pdf/.

European Central Bank (2010). Report on the Lessons Learned from the Financial Crisis. Retrieved April 13th, 2014, from http://www.ecb.europa.eu/pub/pdf/other/reportlessonslearnedfinancialcrisis201004en.pdf.

European Central Bank (2014a). Marking the inauguration of the ECB's new supervisory responsibilities. Keynote speech by Danièle Nouy, Chair of the Supervisory Board of the Single Supervisory Mechanism, at the inauguration of the European Central Bank's new supervisory responsibilities. Retrieved January 29th, 2015, from https://www.bankingsupervision.europa.eu/press/speeches/date/2014/html/se141120.en.html.

European Central Bank (2014b). Regulation (EU) No 468/2014 of the European Central Bank of 16 April 2014 establishing the framework for cooperation within the Single Supervisory Mechanism between the European Central Bank and national competent authorities and with national designated authorities (SSM Framework Regulation) (ECB/2014/17). *Official Journal of the European Union*, L141:1–50.

European Central Bank (2014c). SSM Quarterly Report 2014 / 1 - Progress in the operational implementation of the Single Supervisory Mechanism Regulation. Retrieved February 13th, 2014, from https://www.ecb.europa.eu/pub/pdf/other/ssmqr20141en.pdf.

European Central Bank (2014d). SSM Quarterly Report 2014 / 2 - Progress in the operational implementation of the Single Supervisory Mechanism Regulation. Retrieved May 16th, 2014, from https://www.ecb.europa.eu/pub/pdf/other/ssmqr20142en.pdf.

European Commission (2009). Economic Crisis in Europe: Causes, Consequences and Responses. European Economy 7/2009, DG Economic and Financial Affairs. Retrieved April 13th, 2014, from http://ec.europa.eu/economy_finance/publications/publication15887_en.pdf.

European Commission (2010). Employment in Europe 2010. Retrieved May 7th, 2014, from http://www.eu-bildungspolitik.de/uploads/dokumente_beschaeftigungsfaehigkeit/2010_11_kom_employment_in_europe_2010.pdf.

European Commission (2014). Banking union: restoring financial stability in the Eurozone (press release). Retrieved May 16th, 2014, from http://europa.eu/rapid/press-release_MEMO-14-294_en.htm.

Evans, T. (2010). Five explanations for the international financial crisis. Working Paper No. 08/2010, Institute for International Political Economy Berlin. Retrieved April 13th, 2014, from http://www.ipe-berlin.org/fileadmin/downloads/working_paper/ipe_working_paper_08.pdf.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4):82–89.

Financial Stability Board (2014). Guidance on Supervisory Interaction with Financial Institutions on Risk Culture. Retrieved June 16th, 2014, from http://www.financialstabilityboard.org/publications/140407.pdf.

Groth, S. S. and Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4):680–691.

Gupta, G. K. (2006). *Introduction to Data Mining with Case Studies*. Prentice Hall of India.

Hahn, F. R. (2008). Internationale Bankenkrise und die Rolle von Finanzaufsicht und Finanzinnovationen. *WIFO-Monatsberichte*, 81(8):593–606.

Hájek, P. and Olej, V. (2013). Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. In Iliadis, L., Papadopoulos, H., and Jayne, C., editors, *Engineering Applications of Neural Networks*, volume 384 of *Communications in Computer and Information Science*, pages 1–10. Springer Berlin Heidelberg.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics.

Henselmann, K. and Scherr, E. (2012). Content analysis of XBRL filings as an efficient supplement of bankruptcy prediction? Empirical evidence based on US GAAP annual reports. Working Papers in Accounting Valuation Auditing 2012-2, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Rechnungswesen und Prüfungswesen.

Hille, E. (2013). Recent developments in restructuring the Austrian banking reporting system. *Journal of Banking Regulation*, 14(3-4):269–284.

Hodson, D. and Quaglia, L. (2009). European Perspectives on the Global Financial Crisis: Introduction. *Journal of Common Market Studies*, 47(5):939–953.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University. Retrieved January 21th, 2015, from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Huber, D. and von Pföstl, E. (2013). The Single Supervisory Mechanism within the Banking Union – Novel Features and Implications for Austrian Supervisors and Supervised Entities. In *Financial Stability Report 25*, pages 52–56. Oesterreichische Nationalbank.

Institute of International Finance (2009). Reform in the Financial Services Industry: Strengthening Practices for a More Stable System. Retrieved August 30th, 2014, from http://www.iif.com/download.php?id=ukBoiNBO8UE=.

International Monetary Fund (2010). World Economic Outlook, October 2010: Recovery, Risk, and Rebalancing. Retrieved May 7th, 2014, from http://www.imf.org/external/pubs/ft/weo/2010/02/pdf/text.pdf.

International Monetary Fund (2013a). A Banking Union for the Euro Area: Technical Background Notes. Retrieved May 16th, 2014, from https://www.imf.org/external/pubs/ft/sdn/2013/sdn1301technt.pdf.

International Monetary Fund (2013b). European Union: Publication of Financial Sector Assessment Program Documentation – Technical Note on Deposit Insurance. Retrieved May 19th, 2014, from http://www.imf.org/external/pubs/ft/scr/2013/cr1366.pdf.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg.

Jobst, A. (2008). Back to Basics - What Is Securitization? *Finance and Development*, 45(3):48–49. Retrieved May 11th, 2014, from http://www.imf.org/external/pubs/ft/fandd/2008/09/pdf/basics.pdf.

Kao, A. and Poteet, S. R. (2007). Overview. In Kao, A. and Poteet, S. R., editors, *Natural Language Processing and Text Mining*, pages 1–7. Springer London.

Kapoor, S. (2010). *The financial crisis – causes and cures*. FES Brussels; Bertelsmannstiftung; ETUI, Brussels. Retrieved April 13th, 2014, from http://www.fes-europe.eu/attachments/266_Financial%20Crisis%20-%20causes%20and%20cures%20www.pdf.

KBC Group NV (2011). KBC Annual Report 2010. Retrieved December 17th, 2014, from https://multimediafiles.kbcgroup.eu/ng/published/KBCCOM/PDF/COM_RVG_pdf_08042010_jaarverslag_KBC_Groep_2010_EN.pdf.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33(2014):171–185. Retrieved May 26th, 2014, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2213801.

Kim, K. and Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39(8):1800–1811. Special Issue: Advances of Operations Research in Service Industry.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics.

Korte, J. and Steffen, S. (2014). Zero Risk Contagion - Banks' Sovereign Exposure and Sovereign Risk Spillovers. Working Paper, Goethe University Frankfurt. Retrieved September 20th, 2014, from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2395097`.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season. Working Paper Series. Retrieved January 28th, 2015, from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2408560`.

Leidner, J. L. and Schilder, F. (2010). Hunting for the Black Swan: Risk Mining from Text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 54–59. Association for Computational Linguistics.

Li, F. (2006). Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working Paper, The University of Michigan. Retrieved March 31th, 2014, from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181`.

Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, 48(5):1049–1102.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Liu, B. and Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In Aggarwal, C. C. and Zhai, C. X., editors, *Mining Text Data*, pages 415–463. Springer Science+Business Media.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Lu, H.-M., Chen, H., Chen, T.-J., Hung, M.-W., and Li, S.-H. (2010). Financial Text Mining: Supporting Decision Making Using Web 2.0 Content. *IEEE Intelligent Systems*, 25(2):78–82.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

McCallum, A. and Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, pages 41–48.

Mikut, R. and Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):431–443.

Mishkin, F. and Herbertsson, T. (2006). Financial Stability in Iceland. Technical report, Iceland Chamber of Commerce, Reykjavíc. Retrieved May 7th, 2014, from `http://www.vi.is/files/555877819financial%20stability%20in%20iceland%20screen%20version.pdf`.

Nie, J.-Y., Gao, J., and Cao, G. (2012). Translingual Mining from Text Data. In Aggarwal, C. C. and Zhai, C. X., editors, *Mining Text Data*, pages 323–359. Springer Science+Business Media.

Oesterreichische Nationalbank and Austrian Financial Market Authority (2009). Banking Supervision in Austria. Retrieved March 30th, 2014, from `http://www.fma.gv.at/typo3conf/ext/dam_download/secure.php?u=0&file=3739&t=1378450830&hash=2491ebe0e5edd5c95b54d2a80ddb0997`.

Palley, T. I. (2011). Explaining global financial imbalances: A critique of the saving glut and reserve currency hypotheses. IMK Working Paper 13-2011, IMK at the Hans Boeckler Foundation, Macroeconomic Policy Institute. Retrieved April 13th, 2014, from `http://www.boeckler.de/pdf/p_imk_wp_13_2011.pdf`.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In Shanahan, J., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands.

Power, M., Simon, A., and Tommaso, P. (2013). *Risk Culture in Financial Organisations*. LSE Academic Publishing. Retrieved August 30th, 2014, from `http://www.lse.ac.uk/researchAndExpertise/units/CARR/pdf/Final-Risk-Culture-Report.pdf`.

Quaglia, L., Eastwood, R., and Holmes, P. (2009). The Financial Turmoil and EU Policy Cooperation in 2008. *Journal of Common Market Studies*, 47:63–87.

Robinson, W. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3):351–357.

Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523.

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Sengupta, R. and Emmons, W. R. (2007). What Is Subprime Lending? Federal Reserve Bank of St. Louis Economic Synopses, No. 13, 2007. Retrieved May 8th, 2014, from `https://research.stlouisfed.org/publications/es/07/ES0713.pdf`.

Shull, B. (2010). Too Big to Fail in Financial Crisis: Motives, Countermeasures, and Prospects. Levy Economics Institute Working Paper No. 601. Retrieved January 18th, 2015, from `http://www.levyinstitute.org/pubs/wp_601.pdf`.

Sigurjonsson, T. (2011). Privatization and Deregulation: a Chronology of Events. In Aliber, R. and Zoega, G., editors, *Preludes to the Icelandic Financial Crisis*, pages 26–40. Palgrave Macmillan.

Sprenger, M. (2013). The European Banking Union: Origins, state of play and way forward. Presentation at the Nineteenth Annual Capital Markets Conference of the Federal Reserve Bank of Chicago, retrieved May 22th, 2014, from `https://www.chicagofed.org/~/media/others/events/2013/nineteenth-annual-capital-markets/sprenger-pdf.pdf`.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based Methods for Sentiment Analysis. *Journal of Computational Linguistics*, 37(2):267–307.

Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2152–2157, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thiel, K. and Berthold, M. (2012). The KNIME Text Processing Feature: An Introduction. Technical report, KNIME.com AG. Retrieved December 29th, 2014, from `https://www.knime.org/files/knime_text_processing_introduction_technical_report_120515.pdf`.

Tracy, R. (2014). Bank Execs Try to Push Beyond 'Too Big To Fail'. *The Wall Street Journal*. Retrieved January 18th, 2015, from http://blogs.wsj.com/economics/2014/10/10/bank-execs-try-to-push-beyond-too-big-to-fail/.

Tsai, M.-F. and Wang, C.-J. (2013). Risk Ranking from Financial Reports. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, pages 804–807. Springer-Verlag.

Vourloumis, S. (2012). Reforming EU and Global Financial Regulation: Crisis, Learning and Paradigm Shifts. In *4th Biennial ECPR Standing Group for Regulatory Governance Conference "New Perspectives on Regulation, Governance and Learning 2012"*. University of Exeter.

Voutilainen, A. (2003). Part-of-Speech Tagging. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 219–232. Oxford University Press.

Wooldridge, J. (2002). *Introductory Econometrics: A Modern Approach*. South-Western College Pub, 2nd edition.

Zechner, J., Boot, A., Perraudin, W., Stomper, A., and Zimmermann, H. (2001). Reform of Financial Supervision in Austria - Final Report. Technical report, University of Vienna. Retrieved May 23th, 2014, from http://amor.cms.hu-berlin.de/~stompera/consulting/fma.pdf.