Detecting Risks in the Banking System by Sentiment Analysis

Clemens Nopp
TU Wien
clemens.nopp@alumni.tuwien.ac.at

Allan Hanbury
Institute of Software Technology
and Interactive Systems
TU Wien
hanbury@ifs.tuwien.ac.at

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). While supervisory risk assessments are usually based on quantitative data and surveys, this work explores whether *sentiment analysis* is capable of measuring a bank's attitude and opinions towards risk by analyzing text data. For realizing this study, a collection consisting of more than 500 CEO letters and outlook sections extracted from bank annual reports is built up. Based on these data, two distinct experiments are conducted. The evaluations find promising opportunities, but also limitations for risk sentiment analysis in banking supervision. At the level of individual banks, predictions are relatively inaccurate. In contrast, the analysis of aggregated figures revealed strong and significant correlations between uncertainty or negativity in textual disclosures and the quantitative risk indicator's future evolution. Risk sentiment analysis should therefore rather be used for macroprudential analyses than for assessments of individual banks.

1 Introduction

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 (Hodson and Quaglia, 2009).

As a reaction to the crisis and its severe economic consequences, EU institutions decided to build up an *European Banking Union* (EBU). The EBU consists of three pillars, one of them being a new system of financial supervision, the *Single Supervisory Mechanism* (SSM). Its goal is to "promote long-term 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).

For supervising over 120 of the largest banks in the Eurozone, the SSM utilizes a range of information sources in order to detect vulnerabilities and risks. The sources include mainly backward-looking quantitative *Key Risk Indicators* (KRIs), which are complemented with surveys in order to include forward-looking information as well (European Banking Authority, 2014). 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 this paper, we present a novel application of *sentiment analysis* for exploring attitudes and opinions about risk in textual disclosures by

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 591–600,
Lisbon, Portugal, 17-21 September 2015.

banks. In particular, this work (1) finds suitable data sources, (2) identifies appropriate techniques for risk sentiment analysis, and (3) analyzes risk sentiment within the last decade in order to cover the financial crisis of 2007-08 adequately. The derived sentiment scores quantify uncertainty, negativity, and positivity in the analyzed documents. All of them are interesting with regards to risk sentiment analysis: uncertainty relates to risk in a direct way since the latter are "uncertainties resulting in adverse variations of profitability or in losses" (Bessis, 2002, p. 11). Highly negative sentiment refers to current or future problems, and too positive sentiment could represent overconfidence. We find that sentiment scores reflect not only the financial crisis, but also other major economic events within the last decade.

In addition, we test for correlations between the sentiment scores and a popular quantitative risk indicator. It turns out that aggregated risk sentiment in forward-looking documents is a leading indicator for the actual risk figures, so it can be used within predictive models.

The remainder of this paper, which is based on the Master's thesis of one of the authors (Nopp, 2015), is organized as follows: first, we give an overview on related work in the field of risk sentiment analysis. The following section introduces the chosen sources for text data and quantitative figures. Afterwards, we give an overview on the chosen methodologies and evaluate the experimental results. The last section concludes.

2 Related Work

Sentiment analysis in general and its application in the financial domain in particular gained a lot of interest within the last decade. There is a number of studies which aim to identify risks by means of text mining. 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; 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. Other papers deal with *financial distress prediction*, for example Hajek and Olej (2013). As a baseline, they classify companies based on financial indicators. It turned out that the inclusion of sentiment indicators improved financial distress prediction.

Among the text data sources for these studies are mainly annual reports, but also news stories or earning calls transcripts[1]. Kogan et al. (2009) exclude irrelevant information from the annual reports by focusing on a section which contains important forward-looking content.

In the related studies, authors work with similar approaches for extracting sentiment from texts. Linguistic preprocessing generally involves tokenization, lemmatization, and removing non-essential 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 studies either employ machine learning algorithms, or use the data for regression analyses.

Although none of the mentioned papers focuses on risk sentiment analysis in the banking industry, parts of their processing pipelines and approaches can be reused for this work. With regards to the selection of appropriate data sources, it can be concluded that analyzing annual reports is very popular in this field of research. Hence, these data should also be considered for the experiments of this study. In contrast to the majority of the related papers, we only use specific sections of the annual reports, namely CEO letters and outlook sections (see Section 3).

Regarding the machine learning algorithms and the incorporation of quantitative indicators, the approaches of Groth and Muntermann (2011), Kogan et al. (2009), and Hajek and Olej (2013) are a good basis for the experiments of this study. All of them define the document labels based on suitable quantitative indicators. For labeling, the related studies consider the fact that text data are forward-looking, but quantitative indicators reflect the past. Hence, the indicators are taken from one period after publication of the text data. The labeled data are then used for training machine learning algorithms. Since the focus of our work lies on banks, we make use of a specific quantitative risk indicator which is not employed by related studies. In the following section, we introduce this indicator and the selected text data sources.

[1] Earning calls are regular events where managers report about the company's current situation and answer questions from business analysts.

3 Data Sources

Among this work's aims is to test for relations between textual risk sentiment and quantitative risk indicators. A careful selection of sources for both types of data is crucial since irrelevant ones would lead to biased conclusions.

Quantitative Risk Indicator. The selected quantitative risk indicator has to represent financial health and the general risk exposure of a bank within a specific period or at a specific point in time. Furthermore, the data have to be (1) publicly accessible, (2) available for each analyzed bank, (3) published at least annually, and (4) comparable among the different banks.

A comparison of several quantitative risk indicators based on expert interviews revealed that only the *Tier 1 Capital Ratio* (T1) fulfills all criteria. The T1 is one of the most important ratios based on risk-weighted amount of the bank's assets. In particular, it refers to the bank's Tier 1 capital as a percentage of its risk-weighted assets:

$$\text{Tier 1 Capital Ratio} = \frac{\text{Tier 1 Capital}}{\text{Risk-Weighted Assets}} \quad (1)$$

Tier 1 capital is considered as the best form of bank capital 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 T1 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.

The T1 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 large EU banks in different macroeconomic scenarios, measured by the impact on the T1.

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. Like the quantitative risk indicators, they need to be (1) publicly accessible, (2) available for every analyzed bank, and (3) published at least annually. In addition, for this study, the documents need to be (4) written in the English language, (5) directly published by the bank, and (6) contain forward-looking and subjective information about the bank's attitude and expectations towards risk.

These criteria are best fulfilled by two types of document published in the banks' annual reports, namely *CEO letters* and *outlook sections*. The former are carefully crafted documents which 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, and CEO letters are a way for communicating their attitudes and values.

Outlook sections are usually a part of the *management report*, which is a textual summary of the bank's results, its business environment, and regulatory as well as internal developments. In their outlook on the next year, banks write about the expected macroeconomic environment, management guidelines, and priorities for the next period. These documents might be less subjective compared to CEO letters, but they are usually more comprehensive and contain interesting forward-looking information.

Collection of Data. The annual reports for this work were collected via a *Bloomberg Terminal*, supplemented by direct downloads from bank websites. In total, over 500 documents from 27 banks which published them between 2001 and 2013 were collected. The sample contains banks from all 12 countries which have belonged to the Eurozone at least since 2002. This promotes comparability of the data because the banks operated in similar economic circumstances and with the same currency.

Further data were retrieved from the online database *Bankscope*: the bank's country of residence, its full name, its size measured by total assets, and its Tier 1 capital ratio at the end of each year between 2001 and 2013. 31 % of the Tier 1 capital ratios could not be directly retrieved from the database, so they had to be manually extracted from the respective annual reports.

4 Methodologies

Two independent approaches are employed for the risk sentiment analysis. First, a lexicon-based approach derives and analyzes negativity, positivity, and uncertainty in publications by banks. The second approach aims to predict the evolution of

quantitative risk indicators by means of supervised classification. The aim of both approaches is to assess the potential of risk sentiment analysis in banking supervision.

Creation of the Document Collection. The original documents are provided as PDF files. For building up the collection, they have to be parsed in order to acquire plain text files containing the required sections. One method for extracting the relevant sections is to split the original PDF files according to their bookmarks and to convert them into plain text files afterwards. Another way is to convert the PDF files already in the first step and to extract the relevant sections by making use of specific *tokens*. For example, a typical CEO letter is delimited by the tokens *Dear Shareholders* and *Sincerely*. If neither of these semi-automated approaches is applicable, the extraction has to be done manually[2].

Table 4 gives an overview of the number of documents in the created collection. It shows that the number of published outlook sections constantly increased between 2002 and 2008. From 2009 on, the number was quite stable. The number of CEO letters also increased over time, but only until 2008, when some CEOs stopped writing letters in the course of the financial crisis.

Year	# of CEO letters	# of outlooks
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 1: An overview of the document collection.

4.1 Lexicon-based Approach

The first experiment is about analyzing sentiment scores derived from the documents by incorporating finance-specific word lists. The objective of this experiment is to show how the language of forward-looking disclosures by European banks evolved within the last decade. The workflow consists of the following steps: (1) pre-processing the collected data, (2) the actual sentiment analysis which derives the scores, (3) data consolidation, and (4) data evaluation.

Sentiment Tagging. In the first step of the analysis, sentiment words in the textual data are tagged. In particular, this study works with negative words (*Fin-Neg*), positive words (*Fin-Pos*), and words related to uncertainty (*Fin-Unc*). All of these word lists are provided by Loughran and McDonald (2011). Such topic-specific word lists are necessary because many words bear a different sentiment if used in a financial context: according to Loughran and McDonald (2011), almost three quarters (73.8 %) of typically negative words 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. Table 2 lists some examples for sentiment words in the financial context.

Positive	efficient, stabilized, vibrant
Negative	closure, postpone, threat
Uncertainty	approximately, might, volatility

Table 2: Examples for opinion words in the financial context (Loughran and McDonald, 2011).

Term Weighting. All terms in a document are normalized by

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

This equation is based on Salton and Buckley (1988) and accounts for documents of different lengths. G_i is the *global weight* of term i and $L_{i,j}$ the *local weight* of term i in document j. An established method for the latter is given by the following formula (Manning and Schütze, 1999, p. 543):

$$L_{i,j} = \begin{cases} 1 + \log(tf_{i,j}) & \text{if } tf_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

[2]This was the case for around 20 % of the documents.

The term frequency is denoted as $tf_{i,j}$. The most popular global weight is the *inverse document frequency* (IDF). In

$$G_i = \log\left(\frac{N}{df_i}\right), \tag{4}$$

the total number of documents is denoted by N, and df_i is the number of documents where term i occurs at least once (Salton and Buckley, 1988).

Valence Shifting. In order to account for negated sentiment words, the simple negation handling algorithm proposed by Polanyi and Zaenen (2006) is implemented. If one of the three direct predecessors of a sentiment word is a negation word[3], its sentiment score will be negated. This is done by assigning -1 to the valence shifter variable v_i of term i. If there is no negation word among the predecessors, v_i is set to 1.

Calculating Sentiment Scores. The document-level sentiment scores are calculated for three sentiment classes, namely uncertainty, positivity, and negativity. In

$$s_{c,j} = \sum_{i \in c} L_{i,j} G_i N_j v_i, \tag{5}$$

the term-level sentiment score is represented by the product of the term weights $L_{i,j}$ and G_i, the normalization factor N_j, and the valence shifter v_i. The document sentiment score $s_{c,j}$ is the sum of the term sentiment scores which belong to the document j and the sentiment class c.

Data Consolidation and Evaluation. After calculating the sentiment scores, the data are filtered and grouped in order to prepare them for the evaluations. In particular, the data are filtered according to specific countries and grouped by year respectively by bank.

4.2 Supervised Classification

For the second experiment, the documents are labeled based on a quantitative risk measure, namely the 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.

[3]The considered negation words are *no*, *not*, *don't*, *never*, *none*, and *neither*.

The experiment consists of three steps: (1) reading and parsing the collected data as well as assigning the class labels, (2) linguistic preprocessing and feature selection, and (3) classifying the data with *Naïve Bayes* (NB) and *Support Vector Machine* (SVM).

Assigning the Class Labels. 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, and vice versa. Hence, appropriate labels for the supervised classification task are *UP* for an increasing T1, and *DOWN* for a decreasing one. We assume that the T1 did not change notably if the difference to the previous year is less than 0.2 percent points[4]. In this case, no class label is assigned.

Preprocessing and Feature Selection. Linguistic preprocessing comprises the removal of punctuation, numbers, single characters, and stop words. The remaining words are converted to lower case. Furthermore, the terms are weighted according to the term weighting strategy presented in Section 4.1.

For feature selection, two approaches are followed. The first one assumes that the sentiment words used in the lexicon-based analysis are the relevant features for this experiment. Hence, all words which do not appear in the first experiment's dictionaries are removed. The second approach utilizes a *Snowball Stemmer* to ensure that different versions of the same word are treated as equal. Its feature selection strategy is based on the concepts of *document frequency* (DF) and *information gain* (IG). For DF, tests showed that a lower bound of 20 documents yields the best results. The objective of the IG measure is to identify those features which have the highest discriminatory power in a classification problem. It measures the impurity of a dataset, i.e. its *entropy*. 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

[4]This affected 17 % of the data points in the sample.

of the original data set (Aggarwal and Zhai, 2012, p. 169).

Classification. The outcome of the previous steps is a set of document vectors with associated class labels. With these data, the classification algorithms *Naïve Bayes* (NB) and *Support Vector Machine* (SVM) are trained. The latter is used in its basic version, i.e. with a linear kernel. The performance measures are determined by employing *10-fold cross validation*, which helps to avoid problems like overfitting.

While Naïve Bayes works without parameters, the linear SVM depends on the parameter C. Its optimal value of 111 was determined by conducting an automated *grid search*.

5 Evaluation of the Experiments

Both experiments aim to capture attitudes and opinions about risk by analyzing CEO letters and outlook sections of Eurozone banks. In this section, conclusions are drawn from the results of the experiments.

5.1 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.

Evolution of Sentiment Over Time. Figure 1 shows how sentiment in CEO letters has been evolving since 2002. The evolution of sentiment in outlook sections is not depicted, but is very similar to that of CEO letters. 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 between uncertainty and negativity scores. Since 2012, uncertainty has been decreasing quite sharply. This can potentially 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.

Do Sentiment Scores Predict Quantitative Risk Measures? A comparison of the T1 average evolution and the corresponding sentiment scores reveals interesting relations, see Figure 2. The correlation coefficients in Table 3 indicate that a higher degree of uncertainty or negativity in the documents is commonly followed by a higher increase of the T1, and vice versa.

It is interesting to analyze the data by a regression model for predicting the T1 evolution. Table 4 shows such a model with *negativity* as the only explaining variable. The coefficients can be interpreted as follows: if the average negativity score rises by one unit, the T1 evolution increases by 0.9963 pp. If negativity is zero, the Tier 1 capital ratio would decrease by the computed intercept, which is -0.502 pp. Both coefficients are statistically significant if a 95 % confidence level is assumed.

About 76 % of the average T1 evolution's *variation* can be explained by the negativity score. A model of similar quality could be constructed by analyzing uncertainty in outlook sections. Hence, sentiment scores can be considered as an additional leading indicator for the future evolution of the Tier 1 capital ratio.

Limitations. A drawback of this regression model is that it cannot model *external shocks* which influence the T1 evolution, but are not ad-

[5] A transcript of this speech is available at `https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html`, accessed April 20th, 2015.

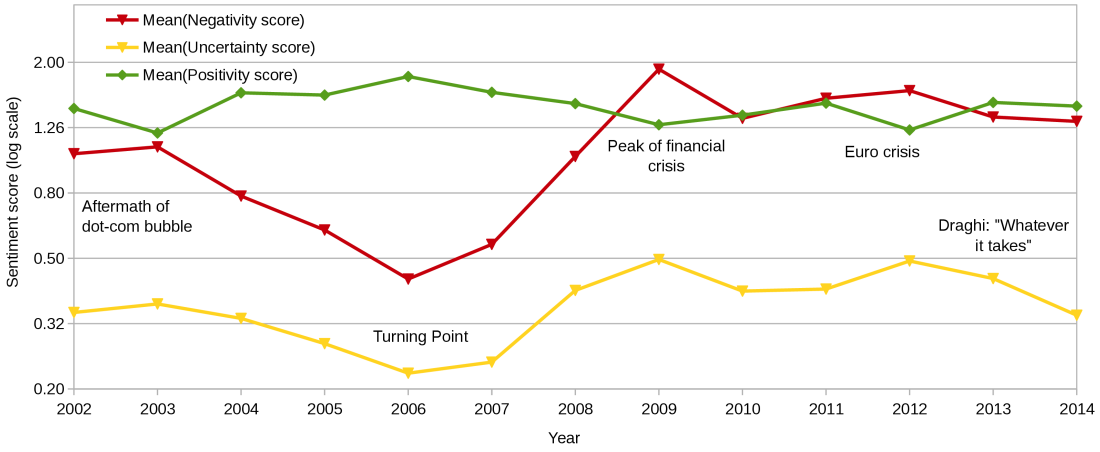

Figure 1: Evolution of positivity, negativity, and uncertainty in CEO letters over time.

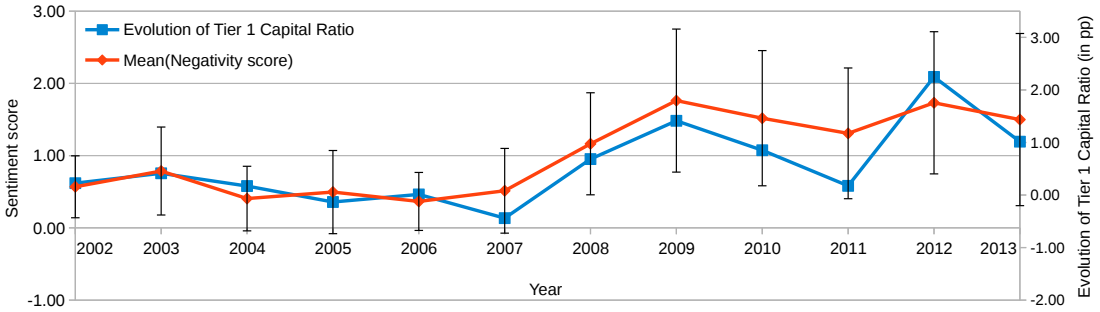

Figure 2: Evolution of the Tier 1 capital ratio compared to negativity in outlook sections. The error bars represent the standard deviation of the negativity scores.

Correlation coefficient	Uncertainty	Negativity	Positivity
T1 evolution (CEO letters)	0.86	0.79	-0.69
T1 evolution (Outlooks)	0.85	0.89	0.12

Table 3: Correlation coefficients between T1 evolution and sentiment scores.

Variable	Coeff.	Std. Err.	t-value	P>t
Mean(Negativity score)	0.9963	0.1647	6.0478	0.0001
Intercept	-0.5020	0.1883	-2.6651	0.0237

Table 4: Regression model based on negativity in outlook sections.

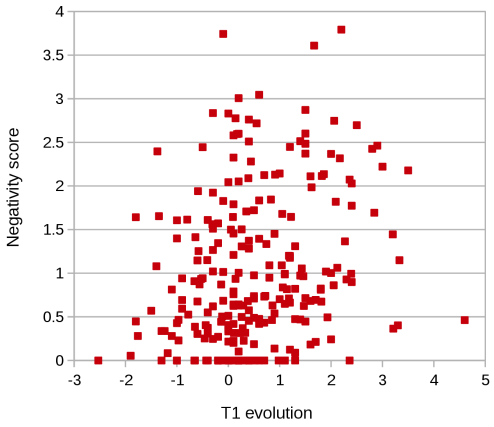

Figure 3: Individual negativity scores in outlook sections compared to the T1 evolution.

equately covered in the text data. Examples for such shocks would be new regulations concerning the minimum capital ratio or monetary policy actions by the ECB. A further limitation is induced by the fact that our methodology makes use of the *bag of words* (BoW) model, which ignores the documents internal structure. Hence, it is not possible to utilize information like word order and grammar, although this definitely plays a role in carefully crafted documents like CEO letters.

It should also be emphasized that the model is based on figures aggregated by year. Applying it on the data of *individual* banks could lead to incorrect conclusions. This assumption is supported by Figure 3, which compares 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 satisfactory representation by a regression model[6]. This observation is in line with the relatively high standard deviations if the figures are aggregated by year, see Figure 2.

5.2 Evaluation of the Supervised Classification Approach

The supervised classification experiment aims to assess whether this approach works better than the lexicon-based approach in terms of predicting the T1 evolution for *individual* banks based on their CEO letters or outlook sections. The class labels *UP* and *DOWN* have been assigned according to the direction of the T1 evolution. Table 5 gives an overview of the experiments and lists the respective performance measures. An analysis of the data in the table reveals interesting results. 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. These classifiers can be seen as a baseline for comparisons. However, the remaining SVM clearly yields the best results among the employed algorithms.

None of the classifiers based on the feature selection method (1) is able to outperform the baseline (assigning every instance to the UP class). Both SVMs simply classify every instance as *UP*, and the Naïve Bayes classifiers also deliver unsatisfactory results. Feature selection based on document frequency and information gain achieves better results than the first one, but only when the classifiers are trained with the CEO letter collection. 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 yields 79.2 %. The other SVM performance measures can be interpreted as follows: 81 % of 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).

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. However, the aggregated classification data accurately predict whether the majority of banks will increase or decrease their Tier 1 capital ratio in the following year: for 12 out of 13 years, the algorithm correctly predicts the direction of the T1 evolution. This finding is in line with the lexicon-based approach, where the aggregated data yielded much better results than the individual ones.

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

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

6 Conclusion

This study explored how banking supervisors could utilize *sentiment analysis* for risk assessments. The analysis of potential document types revealed that two sections in a bank's annual report are particularly well suited for this work, namely *CEO letters* and *outlook sections*. The former 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. Furthermore, 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.

The lexicon-based analysis showed that sentiment scores reflect major economic events between 2002 and 2014 very well. In addition, there is a strong correlation between uncertainty, negativity, and the Tier 1 capital ratio evolution over time. Hence, the sentiment scores could be used in regression models for predicting the T1 evolution. However, the results are only meaningful if the figures are aggregated by year. Applying the model on data of individual banks leads to inaccurate results. 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.

The supervised risk classification approach correctly classifies 79.2 % of the CEO letters. This is not good if one considers that it is possible to yield an accuracy of 70 % simply by assigning the class *UP* to every instance. However, if the results of the best SVM classifier are aggregated by year, the data correctly predict for 12 out of 13 years whether the majority of banks will increase or decrease their Tier 1 capital ratio.

The described systems have the potential to provide valuable insights for banking supervisors, in particular because of the strong correlation between sentiment scores derived from textual data and the T1. Because of the mentioned limitations, these techniques should only be used for macroprudential analyses, i.e. the promotion of stability in the whole financial system. 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.

For future research, 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. The approach could also be used for other document types, for example analyst reports or internal memos, or in other industries. Regarding the methodology, it would be interesting to see how alternative algorithms or word lists would affect the results.

References

Charu C. Aggarwal and Cheng Xiang Zhai. 2012. A Survey of Text Classification Algorithms. In Charu C. Aggarwal and Cheng Xiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer Science+Business Media.

Joel Amernic, Russel Craig, and Dennis Tourish. 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`.

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

Francesco Cannata, Simone Casellina, and Gregorio Guidi. 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.

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.

European Banking Authority. 2014. Risk Assessment of the European Banking System - June 2014, 6. Retrieved July 27th, 2014, from https://www.eba.europa.eu/documents/10180/556730/EBA+Risk+Assessment+Report+June+2014.pdf/.

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

Petr Hájek and Vladimír Olej. 2013. Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. In Lazaros Iliadis, Harris Papadopoulos, and Chrisina Jayne, editors, *Engineering Applications of Neural Networks*, volume 384 of *Communications in Computer and Information Science*, pages 1–10. Springer Berlin Heidelberg.

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

Colm Kearney and Sha Liu. 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.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 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.

Tim Loughran and Bill McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65, 02.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Clemens Nopp. 2015. Risk Sentiment Analysis in Banking Supervision. Master's thesis, Vienna University of Technology.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In JamesG. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands.

Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, August.

Ming-Feng Tsai and Chuan-Ju Wang. 2013. Risk Ranking from Financial Reports. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, pages 804–807. Springer-Verlag.