RatingBot: A Text Mining Based Rating Approach

Diana Hristova
Deutsche Bank AG, diana.hristova@db.com

Johannes Probst
Deutsche Bank AG, Johannes.Probst@db.com

Emanuel Eckrich
Deutsche Bank AG, Emanuel.Eckrich@db.com

Follow this and additional works at: http://aisel.aisnet.org/icis2017

http://aisel.aisnet.org/icis2017/DataScience/Presentations/8

This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICIS 2017 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
RatingBot: A Text Mining Based Rating Approach

Completed Research Paper

Diana Hristova
Deutsche Bank AG¹
Otto-Suhr-Allee 6-16
10585 Berlin, Germany
Diana.Hristova@db.com

Johannes Probst
Deutsche Bank AG¹
Taunusanlage 12
60325 Frankfurt am Main, Germany
Johannes.Probst@db.com

Emanuel Eckrich
Deutsche Bank AG¹
Taunusanlage 12
60325 Frankfurt am Main, Germany
Emanuel.Eckrich@db.com

Abstract

Credit risk is at the core of banking business and its adequate measurement is crucial for financial institutions. Due to lack of historical default data and heterogeneity of customers, qualitative expert-based information is an important factor in measuring the creditworthiness of large companies. However, such information is often extracted manually, causing inefficiencies and possible subjectivity. In this paper we develop the RatingBot: a text mining based rating approach, which efficiently and objectively models relevant qualitative information based on annual reports. It combines both the literature on text mining in finance and machine learning in credit rating to derive the credit rating of a company. We evaluate our approach on two datasets: a publicly available one that facilitates replicability, and a dataset provided by a major European bank representing real-world scenario. The results show that RatingBot delivers additional predictive power and should be considered in future research on credit rating models.

Keywords: Credit risk management, Machine learning, Text mining, Credit rating

Introduction

Risk is at the core of the financial industry as it facilitates the development of different products and services balancing the trade-off between risk-taking and profits. Financial risk management aims at assuring that a company does not take risks excessive of its risk appetite by simultaneously maximizing profits. The most important risk types for the financial industry are credit, market and operational risk (McNeil et al. 2015) where credit risk “…is defined as the risk that a counterparty may become less likely to fulfill its obligation in part or in full on the agreed upon date.” (Christoffersen 2012, p.7). Since one of the main roles of banks is risk transformation by serving as intermediaries between lenders with short-term investment horizon and borrowers seeking long-term financing, credit risk management is at the core of banking business. Its

¹ The views expressed in this paper are those of the authors and in no way represent those of Deutsche Bank AG.
importance is demonstrated by the latest capital figures where for the five major European banks, the average credit risk capital share for 2016 was 70% (Statista 2016).

One of the most prominent parts of credit risk management is its measurement (McNeil et al. 2015), where banks try to reduce information asymmetries regarding the borrowers’ creditworthiness and thus the potential losses in case they fail to repay. Based on the requirements in Basel Committee on Banking Supervision (2001), banks are allowed to develop their own internal rating models for capital calculation. These models aim at quantifying the expected and unexpected loss of a counterparty or a borrower by estimating the three components of probability of default (PD), loss given default (LGD) and exposure at default (EAD) (Bluhm et al. 2016). The PD, being the most crucial to estimate, is commonly mapped to an (internal and external) rating which represents the creditworthiness of the counterparty in the period of one year (Bluhm et al. 2016) and is used for “...loan approval, pricing, monitoring, and loan loss provisioning.” (Grunert et al. 2005, p.2).

In the presence of a reliable historical data, estimating the PD is usually done by considering the historical number of defaults and applying statistical inference (e.g. logistic regression) to quantitative (e.g., interest rate) or categorical (e.g. country) input variables to determine whether a counterparty would default. This is possible for retail customers with large, homogeneous population and many defaults (Lessmann et al. 2013). However, for (large) companies such data rarely exists, as they are characterized by a smaller, heterogeneous population and low default rate. In these cases, a combination of quantitative information, such as a company’s profitability (Grunert et al. 2005), and qualitative, expert-based information, such as quality of management is commonly used (Altman et al. 2010; Godbillon-Camus and Godlewski 2005; Grunert et al. 2005). The reason for using both types of information is that often not all of the factors relevant for the creditworthiness of a company are quantifiable (e.g., future strategy) or publicly available (e.g. some companies disclose limited financial information). However, they can be manually extracted by the credit officer by analyzing different text sources such as extensive corporate disclosures, news, or analyst calls. This qualitative information then adds additional predictive power to the quantitative information (Altman et al. 2010; Godbillon-Camus and Godlewski 2005; Grunert et al. 2005). However, as opposed to quantitative information, the manual extraction process of qualitative information is exposed to inefficiency and subjectivity (Godbillon-Camus and Godlewski 2005) and bears the risk of errors, which are then translated into the PD estimation.

This can have negative consequences in today’s banking environment where banks face pressure from regulators, fintechs, and low interest rates. According to McKinsey currently “...about 50 percent of the function’s staff are dedicated to risk-related operational processes such as credit administration, while 15 percent work in analytics.” (McKinsey & Company 2015, p. 1). With the emergence of Big Data, analytics can be used to derive better credit risk management models, reduce costs, and increase efficiency. One important area of Big Data analytics is text mining which facilitates the automatic structuring and analysis of unstructured text data such as annual reports, news or analyst calls reports.

Many studies examine text data as a source of qualitative information related to the company’s future financial performance. Text sources used in the literature are corporate disclosures, news, and internet postings. Existing works focus on measuring the sentiment of such sources (i.e., “...the degree of positivity or negativity...” (Kearney and Liu 2014, p. 4)) and relating it to the financial performance of a company. Current research has shown that in particular the sentiment of corporate disclosures significantly correlates with future earnings, return on assets, stock returns, and return volatility (Kearney and Liu 2014). However, these works do not consider creditworthiness which is strongly related to the future financial performance of the company and should thus just as well be correlated with textual sentiment. In particular corporate disclosures, such as annual reports, provide well-structured, forward-looking, objective information from company’s insiders on a regular (annual) basis (Kearney and Liu 2014), which makes them suitable for automatically analyzing relevant qualitative information for PD estimation.

The research on credit risk management applying machine learning approaches focuses on classification methods such as neural networks, support vector machines and decision trees. Most of the existing works apply them to a binary classification problem to classify retail customers in ‘good’ (i.e., creditworthy) and ‘bad’ (i.e., not creditworthy) (Louzada et al. 2016) ones. Since, as mentioned above, for such customers

---

2 The value is based on the annual reports of the corresponding banks ordered by total assets.
reliable data exists, this stream of research uses only quantitative or categorical information as input to the classification models and does not consider qualitative information. The few works that consider a non-retail case and qualitative information require a manual extraction process. The purpose of this paper is to combine these two streams of research and at the same time address the efficiency challenges faced by banks nowadays, by developing a text mining based rating approach: the RatingBot. The RatingBot uses as input the annual report of a company and applies text mining and classification approaches to automatically and objectively derive its credit rating. It focuses only on extracting relevant qualitative information, as scoring methods for quantitative information are quite advanced, efficient and well established. It can be used in combination with such scoring methods to develop better credit rating models. Based on the results in the text mining in finance literature, we believe that annual reports are suitable for this task. We demonstrate the contribution of our methodology by applying it to two datasets: 1) a text set of 10-K SEC\textsuperscript{3} (U.S. Securities and Exchange Commission 2017) standardised fillings and corresponding S&P ratings and 2) a text set of non-standardised general annual reports and internal ratings provided by a major European bank. We chose these datasets as 1) allows for a replication of the results and follows the literature where 10-K SEC filings are one of the most commonly used sources (Loughran and McDonald 2016) while 2) represents a real-world scenario with internal ratings. The paper is structured as follows: in the second section we discuss the current state-of-the-art in the two research fields of text mining in finance and machine learning in credit rating and derive the existing research gap. Then we present the RatingBot in the third section, followed by its application to the above two datasets in the fourth one. Finally, in the last section main conclusions are drawn and paths for future research are derived.

Related work

In the following we first present the literature applying text mining in a general financial context to extract quantitative information. This is then followed by a discussion of the literature in the credit risk management field based on machine learning approaches.

Text mining in finance

The literature on text mining in finance is quite extensive and Kearney and Liu (2014) and Loughran and McDonald (2016) both provide very good overviews. In particular Kearney and Liu (2014) discuss the existing works with respect to the source of information, content analysis, and financial modelling. The source of information can be “…public corporate disclosures/filings, media articles and internet messages.” (Kearney and Liu 2014, p. 4). Our focus in this paper is on annual reports where in particular the 10-K SEC filings are a commonly used source in many works (Li 2010; Loughran and McDonald 2011). Content analysis methods, the second criterion considered by Kearney and Liu (2014), aim at extracting relevant information from documents which are in unstructured from. These methods require as input a preprocessed (i.e. cleaned from formatting, punctuations, pictures, tags, etc.,) set of documents and provide as output an interpretable and compressed representation of each document. There are two main approaches in content analysis: bag-of-words and document narrative extraction (Loughran and McDonald 2016). The former assume that the word (or sentence) order is irrelevant for document representation, while the latter derive document representation by semantically and syntactically analyzing its content. Even though the document narrative extraction approaches come closer to the way a human would process a document, existing works in the field are quite rare and still at a very early stage (Loughran and McDonald 2016). Therefore, we focus on bag-of-words approaches here, which can be divided into term-weighting approaches and machine-learning approaches (Kearney and Liu 2014).

Term-weighting approaches structure a document in different terms and assign a weight to each of them representing its importance. There are different ways to calculate the weights and the chosen approach can substantially influence the results (Loughran and McDonald 2016). Many term-weighting approaches use predefined sentiment dictionaries such as Harvard GI word list or the Longhran and McDonald word list (Loughran and McDonald 2011) to determine the sentiment score of a document. The difference between these two word lists is that while the former has been derived in a general context, the latter was extracted

\[\text{https://www.sec.gov/Article/whatwedо.html}\]
from a set of 10-K SEC filings making it more powerful in the financial context (Loughran and McDonald 2011) and thus for credit rating applications. The second type of bag-of-words approaches applies machine learning methods to derive the sentiment score of a document. They require as input a manually labelled set of words or sentences (e.g., ‘positive’) which are then used to train classification model for instance by applying a Naïve Bayes classifier (Li 2010). The model can then be applied for labelling new set of words or sentences. A disadvantage of these approaches is that they are costly to implement, since they require the labelling to be made by native speakers being experts in the finance field (Kearney and Liu 2014) and the amount of data that needs to be labelled (i.e. all terms in all documents) is enormous. As a result, they would not necessarily improve efficiency when applied in the credit risk management context.

Another approach which is sometimes considered in the bag-of-words category and is applied in recent developments in this research stream is topic models which “…identify themes within a corpus of documents.” (Loughran and McDonald 2016, p.27). Here a theme or a topic is defined by a collection of words. Topic models search for latent (hidden) topics that represent the document and thus reduce dimensionality much more effectively than the standard term-weighting approaches. In addition, they deliver much better interpretation in terms of topics, which is crucial for practicable applications. One of the most commonly applied topic models in the literature (Hoberg and Lewis 2017) is the probabilistic Latent Dirichlet Allocation (LDA) (Blei et al. 2003), which assumes a Dirichlet distribution of the topics over all documents and uses Gibbs sampling to derive the topics and their distribution parameters. Topic models may be applied on their own to the preprocessed document or as an additional feature selection step after the term-representation approaches (Sebastiani 2002).

The result of bag-of-words approaches or/and topic models in the literature is often an aggregated quantity (e.g., a sentiment score per document) and is used as input to different econometric models (e.g., linear regression, vector autoregression, or logistic models). These models explain important financial characteristics of a company such as stock prices (Price et al. 2012), trading volumes (Tetlock 2007), future earnings (Li 2010) or fraud (Hoberg and Lewis 2017). However, existing works in this stream of research do not consider the creditworthiness of a counterparty as a dependent variable. In the next subsection, we thus turn to machine learning methods applied in credit rating.

**Machine learning in credit rating**

Most applications of machine learning in credit rating use classification approaches and focus on predicting the default of a retail counterparty by binary classification using only quantitative and categorical information (Khandani et al. 2010; Lessmann et al. 2013). Louzada et al. (2016) identify 187 “…journal papers in English, especially considering ‘credit scoring’ as a keyword related to ‘machine learning’, ‘data mining’, ‘classification’ or ‘statistic’ topics.” (p. 4) in the period 1992-2015. The existing methodologies can be divided into individual classifiers, where only one method is applied, and ensemble classifiers, where multiple homogeneous or heterogeneous methods are considered (Lessmann et al. 2013). Among the individual classifiers, the most commonly applied methods are neural networks (NN), support vector machine (SVM) (Chen and Li 2010), linear and logistic regression (LR), decision trees (DT), fuzzy logic (FL), genetic programming (GP), discriminant analysis (DA), Bayesian networks (BN) and hybrid methods, where the first three were applied in 49.5% of the above 187 papers. Louzada et al. (2016) compare the performance of these approaches and conclude that SVM stand up due to “…high predictive performance and low computational effort…” (p. 28). There are a few papers dealing with the non-retail case, which can be divided into binary and multiclass models. In the former, Altman et al. (2010) apply a LR to predict the default of an SME based on quantitative and qualitative information (e.g., ‘going concern’ audit), while Grunert et al. (2005) use a probit model also on quantitative and qualitative information (e.g., management quality). In both cases the quantitative information extraction requires additional manual effort. In the multiclass case, Moody and Utans (1994) apply NN by mapping the rating classes to the unit interval. Huang et al. (2004) extend a binary SVM by constructing a model for each pair of credit rating classes, while Ahn and Kim (2011) develop a NN model distinguishing one credit rating class against the next one thus reducing the number of models. All three use only quantitative information as model input. Thus, most of the works in this stream of research consider solely quantitative information as a determinant of creditworthiness. The few that incorporate qualitative information rely on inefficient manual information extraction process. Table 1 provides a short overview of this stream of research in terms of input variables, output variable, customer type and accuracy.
Table 1. Machine Learning in Credit Rating Research (except)

<table>
<thead>
<tr>
<th>Source</th>
<th>Input variable</th>
<th>Output variable</th>
<th>Customer</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khandani et al. 2010</td>
<td>52 quantitative customer data</td>
<td>2 classes (non-)delinquency</td>
<td>Retail</td>
<td>≈99.0%</td>
</tr>
<tr>
<td>Louzada et al. 2016 based on literature</td>
<td>14 (AUS)/20 (GER) customer data</td>
<td>2 classes good vs. bad</td>
<td>Retail</td>
<td>73.40%-98.00%</td>
</tr>
<tr>
<td>Altman et al. 2010</td>
<td>22 quantitative and qualitative factors</td>
<td>2 classes (non-)default</td>
<td>SME</td>
<td>75.0%</td>
</tr>
<tr>
<td>Grunert et al. 2005</td>
<td>6 financial ratios 2 qualitative factors</td>
<td>2 classes (non-)default</td>
<td>Corporate</td>
<td>88.9%-88.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>88.3%-89.0%</td>
</tr>
<tr>
<td>Ahn and Kim 2011</td>
<td>14 financial ratios</td>
<td>4 rating classes</td>
<td>Corporate</td>
<td>67.13% to 67.98%</td>
</tr>
<tr>
<td>Huang et al. 2004</td>
<td>21 financial ratios</td>
<td>5 rating classes</td>
<td>Corporate</td>
<td>70.27%-80.75%</td>
</tr>
<tr>
<td>Moody and Utans 1994</td>
<td>10 financial ratios</td>
<td>5 rating classes 16 rating classes</td>
<td>Corporate</td>
<td>63.8% 30.6%</td>
</tr>
</tbody>
</table>

To sum up, on the one hand, the stream of research dealing with text mining in finance does not consider the creditworthiness of a company as a dependent variable. On the other hand, the stream of research in the area of machine learning for credit rating aims at this dependent variable, but without applying text mining to efficiently extract qualitative information and rather focusing on quantitative one. To close this research gap, we combine the two streams by developing the RatingBot in the next section. In particular, we apply term-weighting approaches, topic models and sentiment analysis based on the first stream of research to represent the qualitative information contained in the preprocessed annual reports of companies in a structured way and use this representation as an input to the classification approaches used in the second stream of research to predict the credit rating of a company.

Methodology

The process of text categorization is defined as “...the activity of labelling natural language texts with thematic categories from a predefined set...” (Sebastiani 2002, p.1). In the RatingBot, the natural language texts are the annual reports $a_i, i \in \{1, ..., m\}$ of a company $i$ for $m$ counterparties, while the thematic
categories are the corresponding credit rating classes \( c_i, c_j \in \{1, \ldots, n\} \) for \( n \) classes. We define the set of credit rating classes as a set of integers and not as strings (e.g., AA) to achieve a representation independent of the rating scale. Text categorization consists of two main parts: 1) model development, where the classification model is derived based on preclassified texts, and 2) prediction, where the derived model is applied to new unclassified texts to determine their expected categories (cf. Figure 1). Thus during 1), for a given portfolio, the historical annual reports and known credit rating classes are used to derive the relationship between the textual content and the ratings in the form of the RatingBot classification model. During 2), this model is applied to predict the future credit rating classes of the companies in the portfolio and later use this qualitative information to increase the predictive power of quantitative information. Based on the literature, we first preprocess the preclassified annual reports and then derive a structured and compressed document representation which is finally used as an input to the classification model. In the following subsections, we discuss these three steps. We should note that, even though we concentrate here on annual reports, parts 1) and 2) in Figure 1 can similarly be applied to any kind of text source relevant for companies’ creditworthiness such as news or analyst calls.

**Text preprocessing**

Since annual reports come in a raw, unstructured form, they need to first be cleaned and unified to be able to apply document representation techniques to them (Feldman and Sanger 2007). Step 1 is to clean the pictures, HTML tags, formatting, etc., thereby leaving only the raw text. Then everything is transformed to lower-case letters (Step 2) to avoid double terms with the same meaning (e.g., ‘Capital’ vs. ‘capital’). The text is additionally cleaned by removing numbers, special characters and punctuation (Step 3). Afterwards it is tokenized by removing white spaces and thus separating it into different terms (Step 4). In Step 5 stop words are removed depending on the language. Examples for stop words in English are connectives or articles, e.g. “and”, “or”, “the”. They are “...topic-neutral words...” (Sebastiani 2002, p.13) aiming to make the text more readable for humans, but not really contributing to its content. In Step 6, the remaining terms are stemmed back to their root (e.g., removing the ‘s’ from ‘risks’ or ‘ed’ from ‘declined’), thus grouping terms with the same conceptual meaning together. Stemming algorithms aim to reduce the terms with the same semantical origin to the same root. As opposed to lemmatization, this root must not be the “....linguistically correct root.” (Jivani 2011, p.1931). For example, after stemming, the words ‘goes’ and ‘went’ will not be mapped to the same root, while after lemmatization they will (Jivani 2011). Thus, lemmatization is more precise, but requires a very complex linguistic analysis which usually does not pay off in applications (Hotho et al. 2005). This is the reason why here we only focus on stemming. This completes the description of the preprocessing phase, the result of which is a list of stemmed terms.

**Document representation**

The result of the preprocessing phase contains many terms some of which occur more than once (e.g., ‘revenue’ may appear twice in different parts of the document). To achieve a compressed and interpretable document representation, those occurrences should be aggregated together as they stand for the same meaning. As discussed above, one way to do this is term-weighting (Sebastiani 2002). The simplest term-weighting approach is binary frequency where an indicator function is applied to the occurrence of the term (cf. Table 2). However, this approach ignores the frequency of occurrence of the term, which may be crucial in applications. Thus, a natural extension is to determine the absolute or relative frequency i.e. terms that occur more frequently are assigned higher importance than such that occur less frequently (cf. Table 2). These frequencies capture the information of “...how salient a word is within a given document.” (Manning and Schütze 1999, p. 542). A drawback of the absolute and relative frequencies is that they do not consider the number of annual reports containing a particular term. This is crucial in the context of text classification as, if a term appears with the same frequency in all annual reports, its separation power with respect to the credit rating classes is zero. Thus, a more sophisticated approach to determine the term weights is provided by the term frequency-inverted document frequency (tf-idf) approach (cf. Table 2) which reduces the weight of terms that occur in many documents and increases the one of such that occur in few documents.

The above term-weighting approaches weight every word in the same way, independently of its sentiment. However, the literature has shown a significant relationship between the sentiment of an annual report and the financial performance of a company (cf. second section). Thus, here we also consider a sentiment
weighted term frequency which is based on the word list in Loughran and McDonald (2011) derived in a financial context based on 10-k fillings and thus relevant for credit risk.

<table>
<thead>
<tr>
<th>Table 2. Term Weighting Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Binary frequency</td>
</tr>
<tr>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Relative frequency</td>
</tr>
<tr>
<td>Term frequency-inverted document frequency (tf-idf)</td>
</tr>
</tbody>
</table>
| Sentiment-weighted frequency       | \( Sen_{t,i} = \begin{cases} v_{t,i}, & \text{if } \text{Sen(term)} > 0 \\ -v_{t,i}, & \text{if } \text{Sen(term)} < 0 \\ 0, & \text{else} \end{cases} \) where: 

\( \text{Sen(term)} \) is the sentiment of \( \text{term} \)

\( v_{t,i} \in \{Bin_{t,i}, Abs_{t,i}, Rel_{t,i}, Tf_i d_i\} \)

Table 2. Term Weighting Approaches

Regardless of the chosen term-weighting approach, since annual reports have many pages, generally the number of terms resulting from the above analysis will be very high increasing the probability of overfitting and making the application of classification approaches more difficult and time-consuming. To address this issue, we need to reduce the number of terms by either 1) term selection or 2) term extraction (Sebastiani 2002). The idea behind 1) is to keep only a subset of the initial number of terms and thus increase computational efficiency and the performance of the models, as they can concentrate on the important features. We do this here following the text mining literature by first removing sparse terms i.e. terms that appear very rarely (e.g., in only 5% of the documents). This does not influence the explanatory power as studies have shown that both very rare and very frequent terms do not play an important modelling role (Sebastiani 2002). The next step is to remove terms that provide a low explanatory power with regard to the credit rating classes. To determine these, we apply the chi-square statistic which is commonly used in the text mining literature (Sebastiani 2002).

As opposed to 1) term selection, 2) term extraction aims at deriving a new (smaller) set of terms based on the existing ones without decreasing explanatory power. Here we apply topic models and in particular LDA as this is commonly used in the text mining in finance literature as discussed above. The aim of topic models is a) for each topic \( \text{topic}_h \) to derive the probability distribution \( P(\text{term}_1, ..., \text{term}_n | \text{topic}_h) \), over the terms and b) for each annual report \( a_t \) to determine the probability distribution \( top_{i,h} = P(\text{topic}_h | a_t) \forall h \) over the topics. This is done in LDA by assuming a generative process and using variational methods and the EM algorithm (Blei et al. 2003). For instance, we may have the following two topics in a corpus of annual reports:

- **Topic A**: risk (30%), loss (50%), deteriorating (19%), profit (0%), creditworthiness (1%), improvement (0%)
- **Topic B**: risk (0%), loss (0%), deteriorating (0%), profit (10%), creditworthiness (30%), improvement (60%)

An annual report may have a distribution of 80% Topic A and 20% Topic B. Since the topics are derived only from the terms, the topic probabilities \( top_{i,h} \) then represent term importance equivalently to the weights presented in Table 2. However, they provide much better interpretability for humans. For simplicity from now on we will use \( w_{i,t} \), to represent both the weights in Table 2 and the topic probabilities \( top_{i,h} \) for annual report \( a_t \). This completes the phase of document representation, the result of which are annual

---

4 Generally the terms will be stemmed before applying topic modelling. We present them here before stemming for illustration purposes.
reports in a structured and compressed form. In particular for each annual report \( a_i \) and each term (or topic) \( \text{term}_i \) its weight (or probability) \( w_{i,j} \) is derived. Thus, the result is a structured document-term matrix \( \{w_{i,j}\}_{i\in[1,...,m],j\in[1,...,p]} \) which we use as an input to the machine learning classification approaches in the next subsection.

**Classification**

Classification approaches aim to assign an unclassified instance to one of a set of predefined categories (Han et al. 2011). In our case the instance is represented by the document representation \( \{w_{i,j}\}_{i\in[1,...,m]} \) of the annual report \( a_i \), while the categories are the corresponding credit rating classes \( c_i, c_i \in \{1, ..., n\} \). The idea of such approaches is to train the classifier on a labeled training set with known credit rating classes and later use the developed model to predict the credit rating class of a new annual report. Let, for the rest of this subsection, \( (a_i, c_i)\in[1,...,m] \) \( \Longleftrightarrow \) \( \{w_{i,j}\}_{i\in[1,...,m]} \), \( c_i \in[1,...,m] \) represent this training set and \( \tilde{a}_i \) be a new annual report with unknown credit rating. The classification approaches we apply here are Naïve Bayes (NB), SVM, NN, DT, LR, DA and Supervised Topic Models (STM). The choice of SVM, NN, DT, LR and DA is based on the credit rating literature, while STM is derived from the text mining literature as discussed above. Finally, we also consider NB due to its simplicity and as a benchmark approach. In the following we shortly present each of these classifiers and discuss their advantages and disadvantages in our application context.

The simplest classification approach is the NB classifier, which classifies the instances according to the conditional probability \( p(c_i = k|a_i) \) that a given annual report \( a_i \) belongs to the credit rating class \( k \). It is based on the Bayes’ theorem which states that:

\[
p(c_i = k|a_i) = \frac{p(c_i = k)p(a_i|c_i = k)}{p(a_i)}
\]

and \( a_i \) will be classified in the class \( k \) with the highest \( p(c_i = k|a_i) \). As \( p(a_i) \) is the same for all classes, its value is not relevant for classification. Moreover, based on the maximum likelihood estimator \( p(c_i = k) \) is simply the relative frequency of the credit rating class \( k \) in the labeled data. In order to estimate \( p(a_i|c_i = k) \), NB makes the (strong) assumption that the term weights are conditionally independent from each other and thus:

\[
p(a_i|c_i = k) = \prod_{j=1}^{p} p(w_{i,j}|c_i = k)
\]

Since the term weights described in the previous subsection are continuous variables, here we use the Gaussian form of NB with \( p(w_{i,j}|c_i = k) \sim N(\mu_k, \sigma_k^2) \) where \( \sigma_k^2, \mu_k \) are estimated based on the labelled annual reports using a maximum likelihood estimation. As a result, the classification rule for an annual report \( \tilde{a}_i \) with unknown credit rating class is as follows:

\[
c_{\text{new}}(\tilde{a}_i) = \arg\max_k \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(w_{i,j}-\mu_k)^2}{2\sigma_k^2}\right)
\]

NB classifiers are characterized by good performance, despite their simplicity which makes them popular. However, they may not deliver such a good performance in the presence of many features each with low explanatory power which is typical in text categorization.

The next classification approach we use is SVM. To illustrate it, let us assume that we only have two credit rating classes (i.e., \( c_i \in \{1,2\} \)). The aim of SVM is to find a (linear)\(^5\) hyperplane in the term space that separates those two classes and has a maximal margin \( M \). The annual reports that are closest to the hyperplane are the support vectors and determine the model. The hyperplane is derived by solving the following optimization problem:

---

\(^5\) Here we present for simplicity the linear case with hard margins. In applications, different kernel functions are used to transform non-linear data to linear one (Tong and Koller 2001) and soft margins are applied (Schölkopf and Smola 2002).
\[
\max_{\mathbf{w, b}} \sum_{i=1}^{p} y_i (\mathbf{w}^T \mathbf{x}_i + b) - \frac{1}{2} \|\mathbf{w}\|^2
\]
\[
s.t. (c_i - 1)(\mathbf{w}^T \mathbf{x}_i + b) \geq M \forall i
\]
where \(\mathbf{w} = (b, \ldots, b_p)\). This optimization problem can be transformed to a simpler quadratic optimization problem by applying duality theory (Schölkopf and Smola 2002). The constraints in Equation (4) assure that the classification rule is satisfied for the training set. This classification rule for the unlabeled annual report \(\tilde{a}_i\) is as follows:

\[
c_{\text{new}}(\tilde{a}_i) = \begin{cases} 
1, & \text{if } (\mathbf{w}^T \mathbf{x}_i + b) \geq 0 \\
2, & \text{if } (\mathbf{w}^T \mathbf{x}_i + b) < 0
\end{cases}
\]

The approach can be applied to multiple classes by determining multiple hyperplanes (one-against-one or one-against-all approach) and classifying based on majority voting (Hsu and Lin 2002). SVM are quite robust to overfitting (Feldman and Sanger 2007) and thus the dimensionality reduction performed during the document representation phase is not so crucial. However, they suffer from interpretability issues.

Now we turn to NN as a classification approach. NN are based on the idea of mimicking the human brain and consist of three types of layers: input, hidden and output layer. Each of the layers consists of multiple nodes and the layers are connected to each other with different aggregation and activation functions. There is one input and one output layer and there could be multiple hidden layers. In our case, the inputs to the input layer are the document representations of the annual reports. Before they feed into the next (hidden or output) layer, they are aggregated by applying for instance a weighted sum. Then an activation function such as a logistic or sigmoid function is applied to the aggregated value (Han et al. 2011). This process is repeated for each hidden layer and the output layer. In order to train the model very often backpropagation is used which after each run propagates the prediction error from the output layer back through the previous layers and adjusts the aggregation weights accordingly by applying a gradient descent method.

To illustrate the approach, we consider a simple linear NN with one input and output layer and again two credit rating classes (i.e. \(c_i \in \{1,2\}\)). The idea is first to aggregate the values from the input layer and then check, if they are above a certain activation threshold \(\theta\). Let the aggregation function be a weighted sum with weights \(\beta_1, \ldots, \beta_p\) and the activation function be as follows:

\[
g(x) = \begin{cases} 
1, & \text{if } x > \theta \\
0, & \text{else}
\end{cases}
\]

Then the classification rule for an unlabeled annual report is given by:

\[
c_{\text{new}}(\tilde{a}_i) = \begin{cases} 
1, & \text{if } g \left( \sum_{l=1}^{p} \beta_l \tilde{w}_{il} \right) = 0 \\
2, & \text{else}
\end{cases}
\]

For multiple classes, either multiple models are built comparing classes as for SVM or the output layer contains one node for each class representing its probability and majority voting is again applied. NN have a good predictive quality even for noisy data and deal well with rare, unexpected patterns. However, they take longer to train and are known as 'black-box' models making interpretation difficult (Han et al. 2011). Also, overfitting may be an issue.

The next classifier we discuss is DT. A tree consists of a set of nodes, representing the dependent and independent variables, and a set of branches, representing certain decision rules on the values of the independent variables (e.g., term weight is less than three). The first node of a tree is its root and to arrive at the next node one of the branches must be followed, based on the corresponding value in the document representation. This is done until the last layer of nodes is reached (the leaves) which stands for the credit rating class. The most common DT approaches are CHAID, CART and C5.0 (Louzada et al. 2016) which
differ based on the spilling criterion (e.g., chi-squared, Gini coefficient or entropy-based). To illustrate the approach, let \( R_k = \{ w_{il} \in \text{set}_kl, \forall l \in \{1, \ldots, p\} \} \) represent the set of branches for \( a_i \) to follow to arrive at \( c_i = k \) (e.g., \( R_1 = \{ w_{11} < 3, w_{12} > 0, w_{13} \in [5,8] \} \) for \( p = 3 \)). Then the classification rule for the annual report \( \tilde{a}_i \) is:

\[
c_{\text{new}}(\tilde{a}_i) = c_k | \tilde{w}_{il} \in \text{set}_kl \forall l
\]

DT methods are known for their interpretability and efficiency, however in the context of text mining they can quickly explode in size increasing the probability of overfitting (Lessmann et al. 2013).

LR is one of the most popular methods in the credit rating literature. For two credit rating classes (i.e. \( c_i \in \{1,2\} \)), LR estimates the probability for a given class by using a logit-link function and a linear combination of the term weights i.e.,

\[
P \left( c_i = 1 \mid \{w_{il}\}_{l=1}^{p} \right) = \frac{1}{1 + \exp(\beta_0 + \sum_{l=1}^{p} \beta_l w_{il})}
\]

where \( \beta_0 \ldots \beta_p \) are estimated by minimizing the negative log-likelihood function. The decision rule is:

\[
c_{\text{new}}(\tilde{a}_i) = \arg\max_k P \left( c_i = k \mid \{w_{il}\}_{l=1}^{p} \right)
\]

In the case of multiple rating classes, similar to SVM, the one-against-all approach can be used where one class is compared to all others. LR is easy to interpret, however requires uncorrelated independent variables and may also result in overfitting similar to decision trees. In addition, the logit-link function relationship may not hold in reality.

The next classifier we consider is (linear) DA (Fisher 1936). The original aim of Fisher’s DA was to reduce dimensionality while keeping class discriminant information. It linearly projected the data so that the distance between the projected class means is maximized, while the projected within class variance is minimized (Mika et al. 1999). Modern applications use a multivariate Normal distribution as a prior and Bayes’ theorem. However, it can be shown that under certain conditions the two approaches are equivalent (Hamsici and Martinez 2008). In the following we present the latter, more universal approach by assuming a two-class problem (i.e. \( c_i \in \{1,2\} \)) and linear DA. The prior probability is given by:

\[
p(a_i | c_i = k) = p(W_i | c_i = k) = \frac{1}{(2\pi)^p |\Sigma|^\frac{1}{2}} \exp\left( -\frac{1}{2} \left((W_i - \mu_k)\Sigma^{-1}(W_i - \mu_k)\right) \right)
\]

where \( W_i = \{w_{il}\}_{l=1}^{p}, \mu_k, k \in \{1,2\} \) are the corresponding class means and the covariance matrix \( \Sigma \) is:

\[
\Sigma = \frac{|a_i: \forall c_i = 1|}{|a_i|} \sum_{a_i: \forall c_i = 1} (W_i - \mu_1)(W_i - \mu_1)^T + \frac{|a_i: \forall c_i = 2|}{|a_i|} \sum_{a_i: \forall c_i = 2} (W_i - \mu_2)(W_i - \mu_2)^T
\]

The classification rule for the annual report \( \tilde{a}_i \) is based on classifying in the most probable class as follows:

\[
c_{\text{new}}(\tilde{a}_i) = \begin{cases} 1, & \text{if } \log \frac{p(c_i = 1 | \tilde{a}_i)}{p(c_i = 2 | \tilde{a}_i)} > 1 \\ 2, & \text{else} \end{cases}
\]

which is a linear decision rule since the following expression is a linear combination of \( \tilde{w}_i = \{\tilde{w}_{il}\}_{l=1}^{p} \):

\[
\log \frac{p(c_i = 1 | \tilde{a}_i)}{p(c_i = 2 | \tilde{a}_i)} = (\mu_2 - \mu_1)^T \Sigma^{-1} \tilde{W}_i - \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \log \frac{|a_i: \forall c_i = 1|}{|a_i: \forall c_i = 2|}
\]

(Linear) DA reduces dimensionality to \( n - 1 \) decision boundaries for \( n \) classes. However, it may be difficult to interpret for multiple classes and the linear case may not be able to model complex non-linear relationships.

---

6 Since the result of the LR is a probability, some authors apply a threshold instead of majority voting.
Finally, we discuss STM (Mcauliffe and Blei 2008) which are a type of topic models as presented above. However, as opposed to traditional topic models, which are an unsupervised approach, STM are based on the idea that the topic derivation is done conditional on a given response variable which is the credit rating class $c_i$ here. For example, if an annual report has a distribution of 80% Topic A and 20% Topic B as described above, it would be expected to have worse credit rating than an annual report with a distribution of 1% Topic A and 99% Topic B. Similar to traditional topic models, in STM the topic distribution for a given annual report is derived by a generative process which however not only models topics as a distribution over the terms, but also considers the relationship between the topics and the credit rating classes. The latter is done by estimating a linear regression model on the empirical frequencies of the topics. In the following, this generative approach is presented for each annual report $a_i$:

\[
\theta_{a_i} \sim \text{Dir}(\vartheta) \quad \text{Distribution of topics over documents}
\]

\[
z_{a_i} \sim \text{Mult}(\theta_{a_i}), \text{le} \{1, \ldots, p\} \quad \text{Assignment of a topic to a document}
\]

\[
term_{a_i,t} \sim \text{Mult}(\beta_{z_{a_i}}), \text{le} \{1, \ldots, p\} \quad \text{Assignment of a term to a topic}
\]

\[
c_{a_i} | z_{a_i} \sim \text{Mult}(\eta, \sigma^2) \quad \text{Assignment of a document to a class}
\]

The input parameters to the models are $H$ standing for the number of topics and $\vartheta = [\vartheta_1, \ldots, \vartheta_H]$ determining the distribution of the topics over the documents. In addition, the parameters $\beta = [\beta_1, \ldots, \beta_H]$ provide the distribution of the terms under the topics and are initialized and later estimated. Each $\beta_h$ is a vector of size $p$ giving the probabilities for $\text{term}_t, \text{le} \{1, \ldots, p\}$ under topic $h$ (i.e., $P(\text{term}_t|\text{topic}_h)$). The process works as follows for each annual report $a_i$: first, a distribution $\theta_{a_i}$ over the topics for $a_i$ is drawn from a Dirichlet distribution with parameter $\vartheta$. Second, for each $\text{le} \{1, \ldots, p\}$ a topic $z_{a_i,t}$ is assigned by following a multinomial distribution with parameter $\theta_{a_i}$. Third, based on the assigned topic, a term is assigned following again a multinomial distribution with parameter $\beta_{z_{a_i}}$. Finally, the class is determined by a linear regression\(^7\) with coefficients $\eta = (\eta_1, \ldots, \eta_H)$ and covariates consisting of the empirical distribution of the topics in the document $\overline{Z}_{a_i}$ (i.e., $\overline{Z}_{a_i} = \frac{1}{p} \sum_{t=1}^{p} z_{a_i,t}$). Only the parameters $H$ and $\vartheta$ are given, the parameters $\beta$, $\eta$ and $\sigma$ are estimated by the model by an EM-based maximum likelihood estimation.

Thus, we have the following classification rule for the unclassified annual report $\bar{a}_i$:

\[
c_{\text{new}}(\bar{a}_i) = \sum_{h=0}^{H} \eta_h E(\overline{Z}_{a_i} | \vartheta, \beta, \{w_{e_i}\}_{\text{le} \{1, \ldots, p\}})
\]

where $\overline{Z}_{a_i}$ is the random variable describing the empirical distribution of the topics in the annual report $\bar{a}_i$. This is calculated as the average over all terms in $\bar{a}_i$. Since the predicted class is a real number, it is rounded to the closest integer to get the credit rating class.

Since it is based on term extraction, STM provide much better interpretability than the other classification methods as they come closer to the document narrative extraction. However, they are not capable of modelling the classification problem as sophisticated as other methods due to the linear regression assumption.

### Evaluation

#### Sample description and text preprocessing

To demonstrate the applicability and contribution of our approach, we apply the RatingBot to two datasets. Each dataset consists of a corpus of annual reports and credit ratings which are used to train and evaluate the model. The first dataset consists of standardized 10-K fillings downloaded from the database of the U.S. Securities and Exchange Commission (SEC) and the S&P ratings for the corresponding companies. The

\(^7\) Here we present for simplicity the linear regression case. Generally, for an ordinal objective variable, GLMs should be used.
second dataset was provided by a major European bank and consists of the annual reports stored in the bank's database and the corresponding internal credit ratings. We chose those two datasets as the first one contains data that is publicly available and thus facilitates the reproduction of the results, while the second one refers to a real-world scenario. Furthermore, these datasets confirms the applicability of our approach in practical applications. In particular, as mentioned above, classification approaches require a labelled training set on which to be trained. Since credit ratings are used for many purposes within a bank, such sets exist either in internal bank data (Dataset 2) or via the ratings from external rating agencies (Dataset 1). Thus, there is no need to manually determine the ratings for the purpose of the RatingBot in practical applications, confirming its efficiency. Below, both datasets and their preprocessing are described in detail.

### Dataset 1

The first dataset was derived by downloading the available 10-K filings from the SEC EDGAR database for the period 2009 to 2016. The reason for choosing this time span was to avoid the influence of the financial crises which represents a structural brake in the data and also to facilitate comparability with Dataset 2. This resulted in 34972 10-K filings which were then joined with the 17622 available S&P credit ratings in 2016 (9197 unique companies). The join dataset was constructed by taking the latest S&P rating after January, 2009 \(^8\) and matching it with the SEC statements by applying approximate string matching (Navarro 2001) and assuring that the report was issued at most nine months before the credit rating date. This resulted in 1716 data points for 1351 companies. The reasons for having such a high dropout rate are: 1) not all companies rated by S&P have published 10-k filings (e.g. 8451 of the data points were non-US companies), 2) not all companies that submitted 10-k filings are rated by S&P, 3) the filing date and the rating date do not match (e.g. some companies were rated for the last time in 2013, but have a more recent annual report) and 4) there are uncertainties in approximate string matching. In addition, 228 of the 1716 data points had a defaulted credit rating and were thus removed from the sample since we are aiming at predicting the credit rating class and not the default of the company, as in the latter there is no uncertainty. The final sample consisted of 1488 data points which were preprocessed following the steps described above.

### Dataset 2

The second dataset was provided by a major European bank and consists of the annual reports and internal credit ratings of companies in the period 2009 to 2016. This dataset contains annual reports partly in the format of 10-K filings. As opposed to Dataset 1, the annual reports are not standardized and are not only in English. The dataset consists of 10435 total annual statements, which were reduced to 5508 after removing read-protected and non-English (Feinerer et al. 2013) documents (4737) as well as defaulted and not rated companies (190 data points). Thus, this dataset represents a realistic situation for a bank willing to apply our methodology and having non-US customers with annual reports in English. Note that an extension to other languages can be done analogously. The corresponding annual reports were additionally preprocessed following the steps described above.

### Document representation

The result of the preprocessing phase for both datasets is a set of documents consisting of a list of stemmed (repeating) terms. In order to determine the document representation, the term-weighting approaches in Table 2\(^9\) were applied by using the Loughran and McDonald (2011) dictionary to determine the sentiment-weighted frequency. As discussed above, this dictionary was derived from 10-k SEC findings making it suitable for the credit rating context. Finally, we used term selection to remove sparse terms appearing in less than 5% of the documents and kept the most important terms based on the chi-squared statistic. Figures 2 and 3 present the top features in the two datasets when applying absolute frequency and sentiment-weighted absolute frequency for 19 classes. Both Figures 2 a) and 3 b) show features that are very relevant for credit risk rating such as *illeg, loss, risk, invest, capit, credit, secur, collat, guranti, deriv* as well as some industry-related terms such as *energi or pharmaceut* and geographical terms such as *gree.*

---

\(^8\) For counterparties rated only once before January, 2009 the available rating was taken.

\(^9\) We omit binary frequency as it is inferior to other methods and due to space limitations.
which also could be relevant for the creditworthiness of the company as location and industry and can have an effect on credit risk (Altman et al. 2010). In addition, Figures 2 b) and 3 b) show many negative terms like loss, default, impair, advers, damag, bankruptci, stress which when contained in an annual report would all be a reason for concern regarding the creditworthiness of a company. They also provide many positive terms in this respect such as advanc, benefit, posit, outstand, which would be an indication for an improvement in the creditworthiness of the company. Both figures confirm the relevance of the information contained in annual reports for the credit rating of a company.

![Figure 2. Top Features Dataset 1](image1)

![Figure 3. Top Features Dataset 2](image2)

**Model development and evaluation**

For both datasets, the dependent variable is represented by a set of 19 credit rating classes based on the integer representation of the S&P rating scale and excluding defaults (i.e. $c_i, c_i \in \{1, \ldots, 19\}$ for an annual report $a_i$). The data in Dataset 1 has a bimodal distribution with modes at 10 and at 15, while for Dataset 2 it is unimodal. Moreover, to better examine the explanatory power of the classification models, we additionally introduce another, less granular dependent variable derived from $c_i$, defined as follows:

$$\text{band}_i = \begin{cases} 
1 & \text{if } c_i \leq 7 \\
2 & \text{if } 7 < c_i \leq 10 \\
3 & \text{if } 10 < c_i \leq 13 \\
4 & \text{if } c_i > 13 
\end{cases}$$

(16)
This additional dependent variable represents rating bands and the borders of each band were identified by experts in the field of credit risk management. Finally, to facilitate comparison with the stream of research of machine learning in credit rating, we define a third binary dependent variable $bin_i$ distinguishing between investment ($c_i \leq 10$) and speculative grade companies. The independent variables are the term/topic weights derived in the document representation phase.

We apply all of the presented classification approaches to both datasets by building the model in Python on a 75% training set and evaluating its accuracy on a 25% test set. Since the dependent variables are ordinal, the accuracy stands for the percentage of annual reports that are correctly classified. For topic models, the number of topics was varied to result in the highest accuracy on the test set. The most accurate models per dataset/dependent variable are presented in Table 3. Here, column Dataset1$_{c_i}$ provides the results for Dataset 1 and 19 rating credit classes, while Dataset1$_{band_i}$ provides the results for the same dataset and four rating bands as dependent variable. Finally, column Dataset1$_{bin_i}$ stands for the binary case. Similarly, Dataset2$_{c_i}$ corresponds to Dataset 2 and 19 rating classes, Dataset2$_{band_i}$ represents a dependent variable with four rating bands and Dataset2$_{bin_i}$ stands for investment/speculative grade. In addition, Table 3 provides the weighted average precision, recall, $F_1$-score as well as the one and two notches deviation. The last two measures are very important performance indicators for credit risk models.

### Table 3. Best Model Performance (Test Set)

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset1$_{c_i}$</th>
<th>Dataset2$_{c_i}$</th>
<th>Dataset1$_{band_i}$</th>
<th>Dataset2$_{band_i}$</th>
<th>Dataset1$_{bin_i}$</th>
<th>Dataset2$_{bin_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>DT</td>
<td>SVM</td>
<td>NN</td>
<td>SVM</td>
<td>NN</td>
<td>SVM</td>
</tr>
<tr>
<td>Avg. Precision</td>
<td>0.17</td>
<td>0.43</td>
<td>0.50</td>
<td>0.48</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>Avg. Recall</td>
<td>0.17</td>
<td>0.25</td>
<td>0.45</td>
<td>0.48</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>Avg. $F_1$-score</td>
<td>0.16</td>
<td>0.41</td>
<td>0.46</td>
<td>0.74</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>1 notch deviation</td>
<td>13.2%</td>
<td>14.9%</td>
<td>14.3%</td>
<td>11.7%</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>2 notches deviation</td>
<td>16.7%</td>
<td>18.2%</td>
<td>37.1%</td>
<td>24.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows that the highest accuracy on Dataset1$_{c_i}$ is 16.9%, while the one on Dataset2$_{c_i}$ is 24.9%. Since the random accuracy for $c_i$ is 1/19=5.2%, in both cases the best results are substantially better than random showing that qualitative information in annual reports plays a significant role for the prediction of credit ratings. The higher accuracy on Dataset 2 is attributed to the fact that we have more data points there, allowing for a better modelling. If we turn to the results for $band_i$, we can see that the highest accuracy on Dataset1$_{band_i}$ is 44.9%, while the one on Dataset2$_{band_i}$ is 48.3%. Since the random accuracy here is 1/4=25%, we are again significantly better than random classification. Similarly, for $bin_i$, the highest accuracy for Dataset 1 is 75.5% while on Dataset 2 this is 71.9%. In both cases it is much better than the random 1/2=50%. The best performance for Dataset 2 is delivered by SVM, while for Dataset 1 this is done by DT and NN. This in line with the literature on machine learning in credit rating (cf. discussion above) where SVM is considered the best classifier and NN one of the most commonly used methods. The high performance of DT is surprising as they rarely tend to be as good as SVM. This could be due to the sentiment weighting (cf. Table 4), resulting in less, but more informative terms.

Generally, for a two-class problem, high precision implies that few predictions within a certain class were wrong while high recall implies that many of the true class labels are correctly predicted. The $F_1$ score can be considered as a weighted harmonic mean of precision and recall. In Table 3 we can see that for Dataset 1 and 19 classes, we have both a low precision and recall. The reason for this is that for 6 out of 19 classes the model has a true positive rate (TP) of zero. This is an indication that on this dataset probably classes should be aggregated together for a better model. For Dataset 2 and 19 classes, the precision is higher, while the recall remains quite low. Here we have two classes with a TP of zero implying class aggregation. However, we also have three classes with a precision of one, but a low recall showing that the predictions in these classes were all correct, however a small proportion of the true such rating in the data. Most of the
wrongly classified companies were classified in the mode of the distribution, revealing a problem SVM may have with unimodal data. The precision and recall results for the rating bands and the binary case for both datasets are similar corresponding to the accuracy.

Table 4 presents the detailed accuracy on the test set for both datasets, different term-weighting approaches, band and c_i dependent variables, and the classification methods described above. The entries in the ‘Document representation’ column containing ‘+Sent’ are the models with sentiment-weighted frequency. The bold entries in Table 4 represent the cases from Table 3. The results show that the effect of considering sentiment of the terms depends on the classifier, the dataset, and the type of dependent variable. For NN, the performance almost always reduces when considering sentiment. The reason is that NN are very data-intensive classifiers and reducing the features is not enough to compensate the information gain by the sentiment. For DT, sentiment generally improves performance (cf. above) and thus behaves the opposite to NN. For NB and 19 rating classes, considering sentiment increases performance, while the opposite is almost always true for four rating bands. Thus, it behaves in between DT and NN. Another interesting observation is that while sentiment improves the performance of DA on Dataset 1 and decreases it on Dataset 2, we see the opposite behavior for SVM and the rating bands. However, in the case of DA the differences are within 1%, so could be due the sampling. For SVM, this may be due to the fact that Dataset 1 is written in a slightly more neutral language than Dataset 2 and thus less features are left when considering sentiment. Thus, a similar effect as for NN appears. Finally, applying relative or tf-idf frequency leads to different results than absolute frequency showing that the length and the term occurrences among all documents plays a role, especially as the best results are almost never the ones based on absolute frequency.

Table 4. Model Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Document representation</th>
<th>Accuracy on test set</th>
<th>Dataset1_c_i</th>
<th>Dataset2_c_i</th>
<th>Dataset1_band_i</th>
<th>Dataset2_band_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB (Gauss)</td>
<td>Absolute frequency</td>
<td>4.8%</td>
<td>1.8%</td>
<td>36.8%</td>
<td>31.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency</td>
<td>2.2%</td>
<td>2.3%</td>
<td>40.1%</td>
<td>32.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>4.8%</td>
<td>1.7%</td>
<td>36.8%</td>
<td>31.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Absolute frequency+Sent</td>
<td>5.7%</td>
<td>2.5%</td>
<td>34.4%</td>
<td>31.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency+Sent</td>
<td>8.3%</td>
<td>5.1%</td>
<td>35.5%</td>
<td>38.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tf-idf+Sent</td>
<td>5.7%</td>
<td>2.5%</td>
<td>34.4%</td>
<td>31.5%</td>
<td></td>
</tr>
<tr>
<td>SVM (RBF kernel)</td>
<td>Absolute frequency</td>
<td>15.3%</td>
<td>16.1%</td>
<td>38.2%</td>
<td>35.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency</td>
<td>11.6%</td>
<td>11.6%</td>
<td>32.3%</td>
<td>31.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>13.4%</td>
<td>12.4%</td>
<td>39.3%</td>
<td>42.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Absolute frequency+Sent</td>
<td>15.3%</td>
<td>18.4%</td>
<td>37.4%</td>
<td>36.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency+Sent</td>
<td>11.6%</td>
<td>11.6%</td>
<td>32.3%</td>
<td>31.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tf-idf+Sent</td>
<td>12.9%</td>
<td>24.9%</td>
<td>36.8%</td>
<td>48.3%</td>
<td></td>
</tr>
<tr>
<td>NN (multi-layer)</td>
<td>Absolute frequency</td>
<td>10.8%</td>
<td>10.4%</td>
<td>36.3%</td>
<td>37.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency</td>
<td>11.6%</td>
<td>12.6%</td>
<td>40.6%</td>
<td>40.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>11.3%</td>
<td>11.7%</td>
<td>44.9%</td>
<td>37.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Absolute frequency+Sent</td>
<td>7.3%</td>
<td>11.4%</td>
<td>29.3%</td>
<td>22.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative frequency+Sent</td>
<td>11.6%</td>
<td>11.6%</td>
<td>32.3%</td>
<td>31.1%</td>
<td></td>
</tr>
</tbody>
</table>

10 We omit the binary case due to space limitations.
11 For Dataset 1.8% of the terms are not neutral compared to 10% for Dataset 2.
Table 4 also shows that the worst performance is provided by NB and STM. While for the former this was expected due to its simplicity, for the latter it is quite surprising. One reason could be the use of a linear regression instead of a GLM model to predict the rating class. We expect that modifying the approach in this component would improve performance. Since STM provides better interpretability than the other black-box approaches (i.e. the topics can be easily visualized), it should be kept in the list of possible approaches and improved in the future.

In particular, being able to explain the results of a credit rating model to credit officers and regulators is crucial for its acceptance in practical applications and is not provided by black-box methods such as NN and non-linear SVM.

To be able to judge the performance of our model, we compare the best results derived by it with the results from the literature on machine learning in credit rating in Table 1. Generally, such comparison should be considered with caution as the datasets and the input variables differ. However, it gives an idea of the performance of our model compared to other approaches in the literature. The most common class variable in Table 1 is a binary (non-)default variable. As we do not have default observations in our data, we cannot conduct a direct comparison. However, we can still compare the results of the (non-)investment grade variable as representing ‘good’ and ‘bad’ companies. According to it, our results are in line with those in the literature and comparable to Altman et al. (2010). In addition, our results on the rating bands are worse than Ahn and Kim (2011), but we should consider here that, as opposed to the authors, we are only using qualitative information as input. Finally, our results on the 19 rating classes are a bit worse than the one on the 16 rating classes by Moody and Utans (1994), who however use less classes and only financial information as model input.

To sum up, the evaluation shows that in the best case we are substantially better than random, which implies that qualitative information contained in annual reports has significant explanatory power for credit ratings. In addition, the best performance on the more standardized and neutral Dataset 1 is shown
by DT and NN, while for the bigger, but less standardized Dataset 2, SVM always deliver the best results. The worst performance is provided by NB and STM. Moreover, sentiment-based weighting can lead to less features which however are more informative, thus improving the performance of simpler classifiers such as DT and reducing the one of data-intensive classifiers such as NN. In addition, normalizing the absolute term frequency for the document length and the inverted term frequency generally leads to better results. Finally, our results are in line with the literature presented in Table 1.

Conclusion

In this paper, we present the RatingBot: a text mining rating based approach, which provides an efficient way for considering qualitative information when deriving the creditworthiness of a company. This is of particular importance when measuring credit risk of (large) companies, characterized by smaller, heterogeneous population and low default rate, where qualitative information is commonly used to increase the predictive power of quantitative information. Our approach uses annual reports as a reliable, forward-looking and objective source of textual qualitative information. It combines both the stream of research of text mining in finance and the one of machine learning in credit rating by closing an important research gap between them.

The RatingBot applies the steps of preprocessing, document representation and classification to annual reports. During preprocessing, the annual reports are cleaned and subsequently structured. Then, the resulting terms are filtered and compactly represented during the document representation phase. This is done by grouping them together and representing their importance by their frequency, which can also be weighted with the sentiment of the corresponding term. In addition, the number of terms is reduced by term selection and term extraction. Finally, classification models are trained on the document representation as independent variables and the known credit ratings as dependent variables. The resulting model can then be used to efficiently predict the credit rating of new companies based on their annual reports and combined with quantitative information for optimal results.

We evaluate the RatingBot by applying it to two datasets. The first dataset consist of publicly available data, using 10-K SEC filings and S&P credit ratings and thus assuring replicability. The second dataset was provided by a major European bank and is representative for a real-world scenario. The results show that considering annual reports reflects relevant qualitative information with regard to the credit ratings and thus demonstrate the contribution of our approach. The best performance is delivered by SVM, DT and NN classifiers, while the worst one is given by NB and STM. Moreover, sentiment plays a positive role for simple classifiers such as NB and many classes as well as for DT and a negative one for more data-intensive classifiers such as NN. However, the neutrality of the language can influence this role. Finally, care should be taken with regard to the chosen term-weighting approach, but absolute frequency rarely delivers the best performance. Our contribution thus consists in both combining different streams of research to solve a very relevant problem in finance and comparing the performance of document representation and classification approaches in this context. Our results are in line with the literature on machine learning in credit rating.

Our approach has some limitations which should be addressed in future research. One important question is whether companies can manipulate their annual reports in order to achieve better credit ratings. Thus, in future the model can be adjusted to incorporate a time component capturing such a systematic behavior. Moreover, we apply the seven most common classification approaches. Thus, other machine learning approaches such as deep learning methods (Kearney and Liu 2014), ensemble classifiers (Lessmann et al. 2013), or more sophisticated STM such as GLM (Mcauliffe and Blei 2008) may improve performance. In addition, we concentrate on only one source of qualitative information (i.e., annual reports), but other sources such as news, social media or other internet postings may also be relevant for credit ratings and considering them in the model could be done similarly by applying the steps of preprocessing, document representation and classification. Finally, document narrative approaches such as Part-Of-Speech-Tagging can be applied to address the independence assumption of bag-of-words approaches and provide more interpretable results and other languages such as German could be considered.

References


Blei Committee on Banking Supervision 2001. The Internal Ratings-Based Approach.


