University of Twente

Master Thesis

Estimating Creditworthiness using Uncertain Online Data

Author:
Maurice Bolhuis

Committee:
Dr. ir. Maurice van Keulen
Dr. ir. Djoerd Hiemstra
MSc. Ruben Bos
MSc. Luuk Peters

Database Group
Faculty of Electrical Engineering, Mathematics and Computer Science

October 13, 2015

UNIVERSITY OF TWENTE.

Abstract

The rules for credit lenders have become stricter since the financial crisis of 2007-2008. As a consequence, it has become more difficult for companies to obtain a loan. Many people and companies leave a trail of information about themselves on the Internet. Searching and extracting this information is accompanied with uncertainty. In this research, we study whether this uncertain online information can be used as an alternative or extra indicator for estimating a company's creditworthiness and how accounting for information uncertainty impacts the prediction performance.

A data set consisting 3579 corporate ratings has been constructed using the data of an external data provider. Based on the results of a survey, a literature study and information availability tests, LinkedIn accounts of company owners, corporate Twitter accounts and corporate Facebook accounts were chosen as an information source for extracting indicators. In total, the Twitter and Facebook accounts of 387 companies and 436 corresponding LinkedIn owner accounts of this data set were manually searched. Information was harvested from these sources and several indicators have been derived from the harvested information.

Two experiments were performed with this data. In the first experiment, a Naive Bayes, J48, Random Forest and Support Vector Machine classifier was trained and tested using solely these Internet features. A comparison of their accuracy to the 31% accuracy of the ZeroR classifier, which as a rule always predicts the most occurring target class, showed that none of the models performed statistically better. In a second experiment, it was tested whether combining Internet features with financial data increases the accuracy. A financial data mining model was created that approximates the rating model of the ratings in our data set and that uses the same financial data as the rating model. The two best performing financial models were built using the Random Forest and J48 classifiers with an accuracy of 68% and 63% respectively. Adding Internet features to these models gave mixed results with a significant decrease and an insignificant increase respectively.

An experimental setup for testing how incorporating uncertainty affects the prediction accuracy of our model is explained. As part of this setup, a search system is described to find candidate results of online information related to a subject and to classify the degree of uncertainty of this online information. It is illustrated how uncertainty can be incorporated into the data mining process.

Acknowledgements

This master thesis is the result of the graduation of the study Computer Science at the University of Twente and was completed as part of an internship at Topicus Finance. I would like to thank Maurice for his guidance and continuous feedback. I would like to thank Djoerd for reviewing my thesis. Furthermore, I would like to thank everyone at Topicus Finance who supported me during this research. In particular, I would like to thank my supervisors at Topicus Finance, Ruben and Luuk, for their continuing support and valuable feedback.

Maurice Bolhuis
Enschede, October 2015

Contents

Chapter 1

Introduction

1.1 Motivation

Since the financial crisis of 2007-2008, rules for credit lenders have become stricter. Banks are required to maintain larger capital buffers and reduce the capital leverage. As a result, it has become more difficult for companies to obtain credit.

Currently, when a company owner applies for a business loan, the decision by the lender of whether or not to grant the loan is based on two main factors. The first factor is based on the company figures the credit applicant supplies. These figures are supplied in the form of a balance of payments and income statement. However, figures do not paint a complete image of the company. The second factor that influences the decision whether or not a loan is granted, and which maybe even more important is qualitative data. When the Bank and Currency Committee asked J.P. Morgan in 1912 if a man's money and property was the most important factor in lending money, he responded by saying: "No, sir; the first thing is character." [1, 2]. What this example illustrates is that, although a credit applicant might be financially healthy, a credit lender is not likely to lend its money if it does not know if it can trust the applicant. Credit lenders obtain this qualitative data via a conversation with the owner of the company about the financing requirement. This is intended to obtain a better insight in who manages the company and how the company is managed. During this meeting issues like management structure, management capacities and business planning are discussed.

A credit rating score can be calculated over both these quantitative and qualitative factors. A credit rating expresses the probability of default (PD) of a company within a certain period, usually one year. This rating can either be expressed as a percentage indicating the probability of default, or as a rating class which is a combination of

numbers and/or letters. A credit score can influence a lender's decision to grant the credit, but also the terms of the credit, such as the percentage rate, repayment period and additional fees [3]. An increased accuracy of credit scores might broaden the amount of companies that successfully obtain a credit, because it reduces the uncertainty for the lender which is therefore more likely to lend money.

1.2 Problem statement

Many people and companies leave a trail of information about themselves on the Internet. Social media is becoming an increasingly important channel to gather insights on company performance [4]. Data generated from online communication acts as "potential gold mines for discovering knowledge", according to Dey and Hague [5]. As the statement of J.P. Morgan in 1912 already showed, especially qualitative data about a person is of interest to a lender. This information might be found on social media and other websites on the Internet. The financial crisis showed the importance of being able to make an accurate prediction of the creditworthiness of credit applicants. Searching and extracting this information is accompanied with uncertainty. For example, it might be uncertain if the information found is actually related to the person or company that was searched for. Topicus Finance is interested in whether this online information can be used as an alternative or extra indicator for whether or not to grant a loan to a company and how accounting for information uncertainty impacts the performance of this indicator. Figure 1.1 illustrates this problem in which online information, for which it is uncertain that it belongs to the company of interest, is used as input for a creditworthiness prediction model.

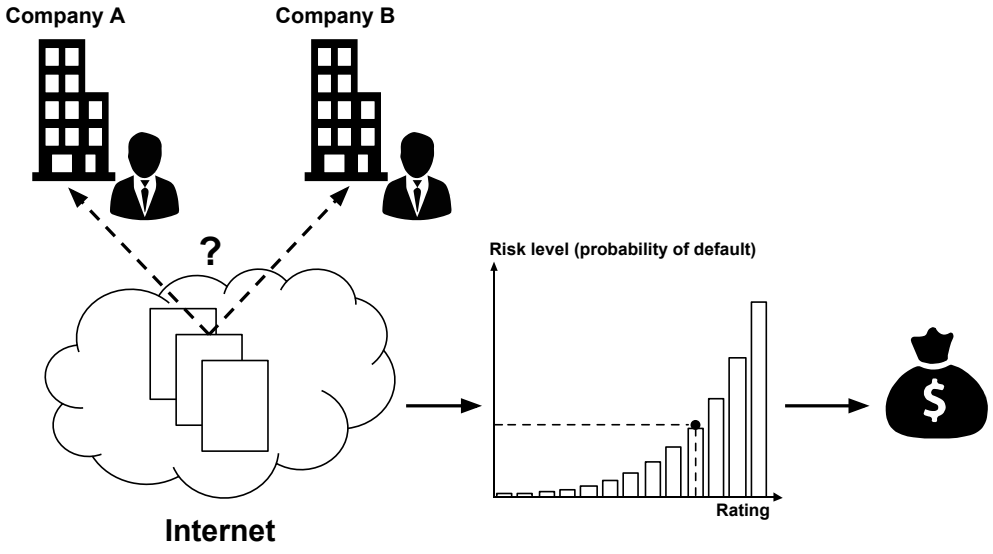

FIGURE 1.1: Online uncertain data as a predictor of the creditworthiness of a company.

In recent years, multiple companies started using information of the Internet for calculating credit scores, such as Lenddo Inc.[1], Neo Finance Inc.[2], Friendly Score Ltd.[3], Affirm Inc.[4] and Kabbage Inc.[5]. Websites used by these companies are mostly social media and include Facebook, Twitter and LinkedIn [6]. Although there are existing solutions, the precise working of their algorithms remains unknown, because it is part of their business model. Most of these existing solutions focus on consumer credit scores, while in this research, we are interested in corporate credit scores. In addition, these existing solutions have in common that they explicitly ask access to a person's social media account. This has several disadvantages. First, it might not always be a viable option to explicitly ask access to a person's social media account, e.g., when information on all the employees of a company would be used for making a creditworthiness prediction. Second, by explicitly asking access to the account, it is more likely that the profile is first fine-tuned before applying for a loan. As a result, the profile data might give an incorrect image of the creditworthiness.

1.3 Research questions

The following main research question is defined based on the problem statement above:

RQ Given a limited and uncertain set of online data about a real world person or company, can a prediction of their creditworthiness be made using this data?

This main research question is divided into the following subquestions:

Q1 Which financial data related to creditworthiness is available to test the performance of online information as predictor for a company's creditworthiness?

Q2 Which online available information can be used as a predictor for the creditworthiness of these companies?

Q3 What is the performance of these indicators as a predictor for creditworthiness?

Q4 How does incorporating the uncertainty of online data affect the prediction accuracy?

[1] https://www.lenddo.com/
[2] https://www.neoverify.com/
[3] http://www.friendlyscore.com/
[4] https://www.affirm.com/
[5] https://www.kabbage.com/

The goal is to build a data mining model that makes creditworthiness predictions based on uncertain online information. Training and testing such a prediction model requires a data set consisting of both financial and online data. Answering the first subquestion helps to acquire the necessary financial data for relating online information to creditworthiness. The second subquestion helps to identify which uncertain online information can be used for predicting the creditworthiness of a company. Answering the third subquestion helps to determine how this online information performs as a predictor of a company's creditworthiness. The fourth subquestion helps to determine whether taking uncertainty into account has a positive effect on prediction accuracy or not.

1.4 Research method

The first research question will be answered by first holding interviews with employees of Topicus Finance about which data related to creditworthiness is available within Topicus Finance or Topicus Finance has/can get access to. The goal of these interviews is to also get clear which steps are required to obtain this data. It might be necessary to create a program to collect the data and construct the financial data set, depending on the availability and completeness of this data.

The second research question will be answered by holding a survey and consulting literature on online information sources that possibly are relevant to creditworthiness predictions of companies. The survey will be held among employees of Topicus Finance. The literature study will help to identify whether there are existing solutions and what kind of information is used within these existing solutions. In order to construct a model, it is required that a significant amount of companies from the financial data set can be found in an information source. Hence, a test will be performed using a sample from this data set to determine the availability of information on the companies within the information sources. Because the model will be trained on data for which uncertainty is not modeled, the accounts of companies and persons related to the company will be manually searched once suitable information sources are found. The eligible information will be harvested from these accounts.

The third research question will be answered by doing two experiments. First, we will test whether solely Internet features can be used for making creditworthiness predictions. Second, we will test whether Internet features can be used for making creditworthiness predictions when used in combination with financial data. For both experiments, data mining models will be constructed using several classifiers. The prediction models will be trained and tested using the financial and online data collected while answering subquestions 1 and 2.

FIGURE 1.2: Steps involved in answering the 4th subquestion.

The research method to answer the fourth subquestion is to perform an experiment in which the model of the third subquestion is reused and in which its accuracy is tested both using data for which uncertainty is and is not modeled. In this research, we focus on uncertainty that arises when searching for online information. Figure 1.2 shows the steps required to answer the fourth subquestion. Literature will be studied on methods of finding online manifestations of persons and companies, such that a search system can be built that can find candidate results. Information of interest will be harvested for these results and will be used for making a creditworthiness prediction using the data mining model. In the certain case, only the information of the best matching search result will be harvested and used. In the uncertain case, the alternative search results are each taken into consideration with their probability. Literature will be studied for methods of incorporating uncertainty into data mining models.

Answering these four subquestion allows to answer the main research question.

1.5 Contributions

The contributions of this research are threefold. First, this research shows online information that is possibly useful for predicting a company's creditworthiness and which is not used in common corporate ratings. Second, it shows how well this information performs as creditworthiness indicator. Third, this research discusses how the uncertainty of online data can be included into the prediction model.

1.6 Report outline

The remainder of this report is organized as follows: Chapter 2 describes the research domain by explaining the typical structure of a credit application process and how risks are estimated within this process. Chapter 3 describes related work. Chapter 4 answers the first research question by explaining how a data set was created that consists of financial data related to creditworthiness. In chapter 5 we answer the second research question by discussing several possible online information sources. Chapter 6 answers the third research question by explaining the experiments performed to test the performance of these information sources as a predictor of a company's creditworthiness. Chapter

7 discusses the experiment for testing how including uncertainty affects the prediction accuracy of the model. Chapter 8 discusses ethical concerns of using online information. Chapter 9 gives the conclusions of this thesis and discusses future work.

Chapter 2

Research Domain

In order to understand how online information can be applied in the credit application process, we explain the steps involved in obtaining a credit and the methods used by credit lenders to estimate the risks. These topics are considered background information for the rest of this report.

2.1 Credit application process

A large fraction of starting entrepreneurs and existing companies that want to invest need to find sources of finance. This financial requirement can be fulfilled using owners' equity or loans. Loans are obtained by consulting a credit lender, often a bank. Before a loan is granted, the credit lender first requires some information about the credit applicant in order to make a deliberate decision. This information is obtained within the credit application process. This is the process in which the credit applicant and credit lender exchange information necessary for the credit lender to make a decision on whether the credit is granted and to determine the amount of credit, the interest rate and the duration of the repayment period. The steps involved within this process are illustrated in Figure 2.1.

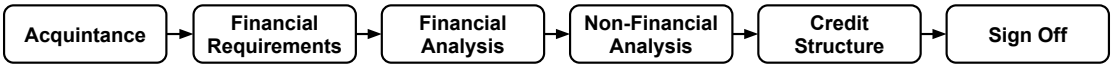

FIGURE 2.1: Steps involved within the credit application process.

The results of the process are written down in a credit proposal. The goal of drafting a credit proposal is to give both the credit applicant and the lender insight into the risks of providing a credit and allowing them to make a deliberate decision. Banks, such as ABN AMRO [7], often follow a step-by-step plan that guides them through the process

of obtaining the necessary information. Although there is no standardized plan, the essence of these steps is roughly the same for different credit lenders.

2.1.1 Acquaintance

The first step is to gather some basic information about the applicant. This includes information such as the type of company of the applicant; whether the applicant is an existing or a new customer; and the reason of his or her visit, i.e., whether the applicant requests a new loan or wants to revise an existing loan. Whenever the applicant is already a customer, it has established a reputation which will be taken into consideration when determining the terms of the credit.

2.1.2 Financial requirements

In this step the credit applicant underpins its financial requirements and the credit lender analyses these requirements. The applicant explains for what it intends to use the credit and motivates the amount of money required, how the credit influences the company, and explains how it intends to repay the loan. The credit lender estimates whether the requested amount of credit is enough for the different business scenarios. It determines which part of the finance can be provided using a credit and which part using equity or third-party resources. Furthermore, they discuss issues like which income sources will be used to repay the loan.

2.1.3 Financial analysis

Within this step the credit lender analyses credit risks by analyzing the cash flows that are expected based on the future prospects of the company. This gives it the credit lender insight into the ability of the company to meet the repayment obligations. The credit lender analyses the financial impact on the company of several future prospects. Most important figures needed for an analysis by the credit lender are supplied using a financial account. A financial account of a company gives a numerical overview of the revenues and costs; assets and liabilities; and income and expenditure of a company. Credit lenders calculate some formulas over these figures to classify and compare credit risk (See Section 2.2).

2.1.4 Non-financial analysis

The financial figures of a company do not paint a complete image of the credit applicant. Despite having good financial figures, a company still can go bankrupt, for example, due to poor management. As explained in Section 1.1, qualitative data, such as the character of an entrepreneur, can be at least as important. This is why credit lenders also analyze non-financial issues. The credit lender considers the management structure of the company, the role of the credit applicant within the management, whether management has ownership interests in the company. They determine whether the company has well-skilled managers that, for example, are capable of dealing with a company crisis situation. Furthermore, it is important that someone within management has experience with financial administration. Besides information about management, the credit lender also requests more general information about the company, such as its activities, the products and services it delivers and the customers it has. They discuss the company's strategic and financial plan.

2.1.5 Credit structure

In this step the credit lender analyses possible structures that impact their risk. Structural risk is the risk that the credit lender has no grip on cash flows [7]. This can arise when other stakeholders come first in line to claim money when the company goes bankrupt. The credit lender analyses which other stakeholders exist and determines its debt position. Furthermore, credit lenders also take into consideration the securities offered by the credit applicant in case the company would default, for example, inventories or sureties.

2.1.6 Sign off

Within this last step the credit lender evaluates the results of all previous performed analyses. It estimates the total risk and underpins why it is justified to accept these risks or not. The lender will determine the interest rate of the loan. It will then consult with the credit applicant and depending on the decision of both a contract is signed to confirm the deal.

2.2 Credit rating

A credit rating expresses the probability of default (PD) of a debtor [8]. It is an assessment of a debtor's creditworthiness represented as a percentage or rating class which

is a combination of numbers and/or letters. A debtor is said to default when "it fails to meet the payment obligations within a specific period" [9]. The period considered is usually one year. A credit rating basically expresses how likely it is that the debtor will repay the debt [10].

Entities for which credit ratings can be calculated vary from entire countries to companies and individual consumers, however, this research only focuses on company ratings. In a guide by the European Commission about ratings [10] a distinction is made between two types of ratings:

- External ratings: these are issued by rating agencies for only countries and relatively large companies.
- Internal ratings: these are assigned by banks to their borrowers.

A rating is calculated using the quantitative and qualitative information acquired during the credit application process. Figure 2.2 illustrates how corporate borrowers are rated.

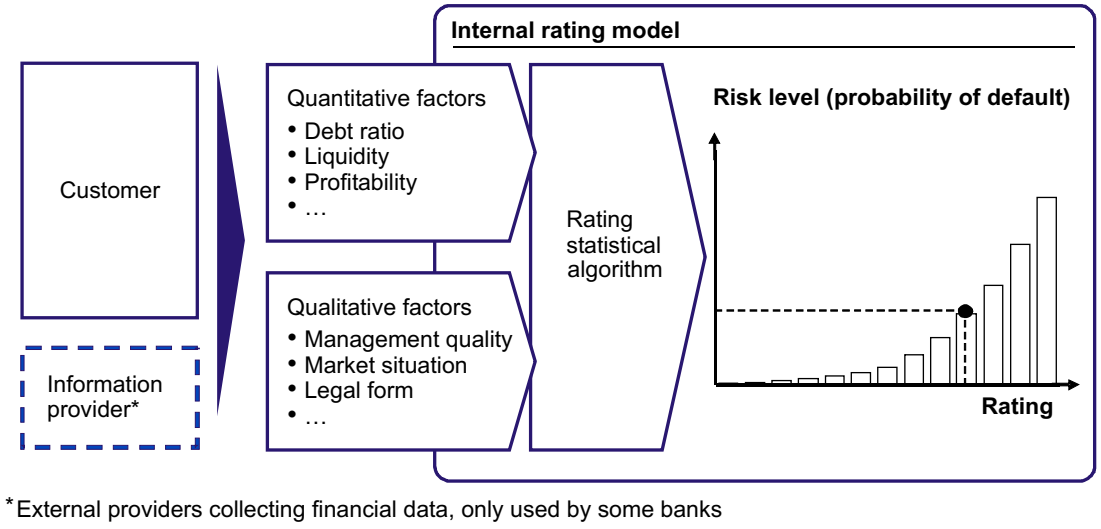

FIGURE 2.2: How borrowers are rated [10].

As explained in Section 2.1, the quantitative data is gathered from financial statements or annual reports. In addition, sometimes the company is also asked for tax returns and business plan figures. Certain ratios are calculated over these figures. Most commonly, this includes ratios that express the indebtedness, liquidity and profitability of the company [10]. Based on the judgment of the bank, values are assigned to the qualitative factors, such as quality of management and market situation. The weight banks assign to these qualitative factors in their rating algorithm often depends on the size of the company and the size of the loan requested [10].

An example of a credit rating score that is considered throughout this thesis is the URA Solvency Check [11]. Although the ratios calculated differ per specific credit rating, Henking et al. [9] show some example ratios that can be used in calculating a credit rating score:

- Share of ordinary results. This ratio takes the income resulting from day-to-day operations of the company as a fraction of the liabilities.
- Liability ratio. This ratio measures the amount due to banks and suppliers as a fraction of the liabilities.
- Interest rate on borrowings. This ratio is calculated by taking the amount of interest expenses as a fraction of the debts. It expresses the estimated level of risk by lenders to the company.
- Share of short-term liabilities. This ratio is calculated by dividing the short-term liabilities by the sales value. A liability is often short-term if it is due within 1 year. This ratio expresses how well the company is able to repay its short-term debts.
- Share of own funds. This ratio measures the fraction of the assets that is financed by owners' equity. A larger ratio shows that shareholders are more willing to invest in the company, therefore making it less risky for potential lenders to invest their money into the company.

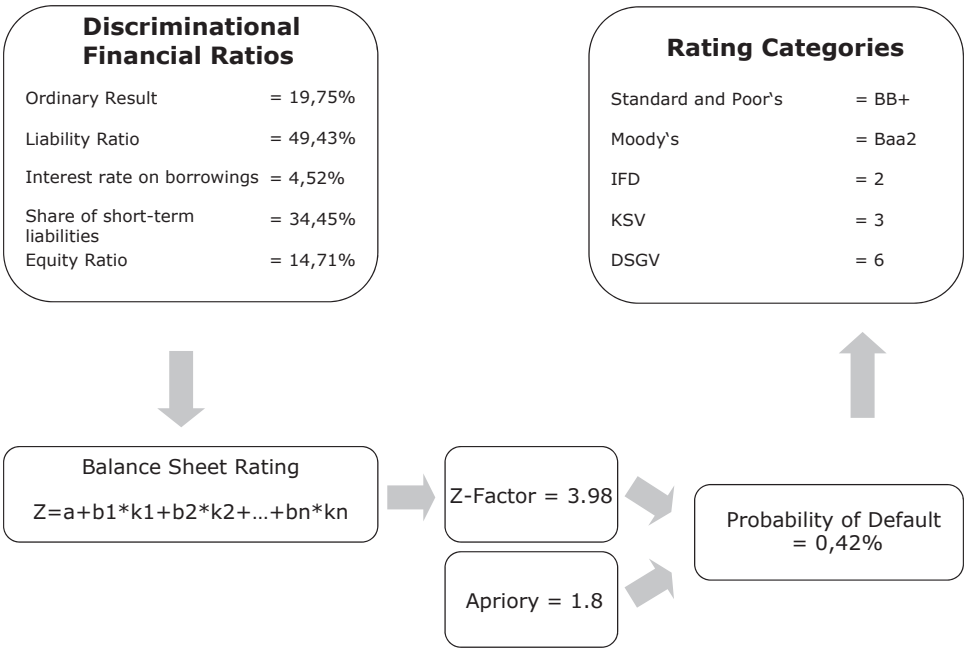

FIGURE 2.3: Example procedure of a rating [12].

Figure 2.3 illustrates the process of calculating a credit rating based on the balance sheet figures. The values of these ratios are inserted into a regression function which transforms them into a single number. This value is normalized using a logistic function into a number between 0% and 100% which indicates the probability that the company goes bankrupt. The company is assigned a particular rating class based on these normalized values. Table 2.1 shows to which rating class a company with a certain probability of default is assigned according to the URA rating.

TABLE 2.1: URA Rating Scale (one-year probability of default) [11].

PD-Rating	**PD%**	**Description**
AAA	0.001%	The company has a very strong financial standing. The ability
AA+	0.002%	to meet its payment obligations is excellent.
AA	0.004%	
AA-	0.008%	The company's solvency is satisfactory. Fundamental strength
A+	0.01%	is not as good as in AAA/ AA+.
A	0.02%	
A-	0.04%	The company's solvency is adequate. However, negative
BBB+	0.09%	changes in the economic conditions may have considerable
BBB	0.17%	effects on the company's ability to meet its payment
BBB-	0.42%	obligations.
BB+	0.87%	The financial standing of the company is characterized by
BB	1.56%	considerable, continuous uncertainty. Distinct speculative
BB-	2.81%	elements. Presently the company is still able to fulfill all
B+	4.68%	payment obligations.
B	7.16%	
B-	11.62%	There is presently a high risk for the company not being able
CCC+	15.40%	to meet its payment obligations on time.
CCC	17.38%	
CCC-	21.50%	Due to its financial situation the company could not meet its
CC	26.00%	financial obligations on time. Selective default in payment,
D	50.99%	insolvency. Failure to pay, insolvency.

Chapter 3

Related Work

Uncertainty can arise when searching for online data. Section 3.1 discusses related work on searching online manifestations of persons. The use of online information for predicting creditworthiness has been studied by related work. Section 3.2 discusses related work in which social media information is used for predicting the creditworthiness of persons. Section 3.3 discusses related work that analyzes the impact of using information of social media connections on customer rating accuracy. Related work on probabilistic databases, the concept of which is applied within this research, is discussed in Section 3.4.

3.1 Online entity resolution

As explained in Section 1.2, searching and extracting online information can be accompanied with uncertainty. For example, uncertainty can arise when resolving online entities. H. Been built a prototype of a system that uses data about a real world person to automatically find online manifestations of that person [13]. The prototype allows finding online manifestations by searching on several characteristics of a person, for example, first name, last name and email address. Possible results are crawled and match scores are calculated for these results using the information known about the person that is searched for. A probabilistic view is applied to derive conclusions from the evidence that is found and calculated over the collected data. Experiments performed with the prototype in which Twitter was used as an online source show that the prototype is able to reliably and automatically find persons on the Internet.

The metrics described by H. Been to determine the level of confidence that a correct search result is found are relevant for this research because they can be used as a measure of uncertainty. Appendix A explains several of these similarity functions that are

discussed in the work of H. Been. It allows determining the best match that can be used for testing the prediction performance when uncertainty is not modeled. When uncertainty is modeled, the metrics can be used to attach a level of uncertainty to the data that is extracted.

In addition to searching information, uncertainty from entity resolution can also arise within data. Related work exists that study entity resolution within informal texts, such as Twitter tweets [14, 15]. Appendix B explains the subtasks of information extraction used by related work. This can be used to determine the level of uncertainty when informal texts are used as an information source for predicting creditworthiness.

3.2 Microcredit risk assessment using crowdsourcing and social networks

Related work has studied the use of online information for predicting creditworthiness. In [16], Hasanov et al. explain a system that performs credit risk assessment using crowdsourcing and social networks. Figure 3.1 illustrates the proposed system.

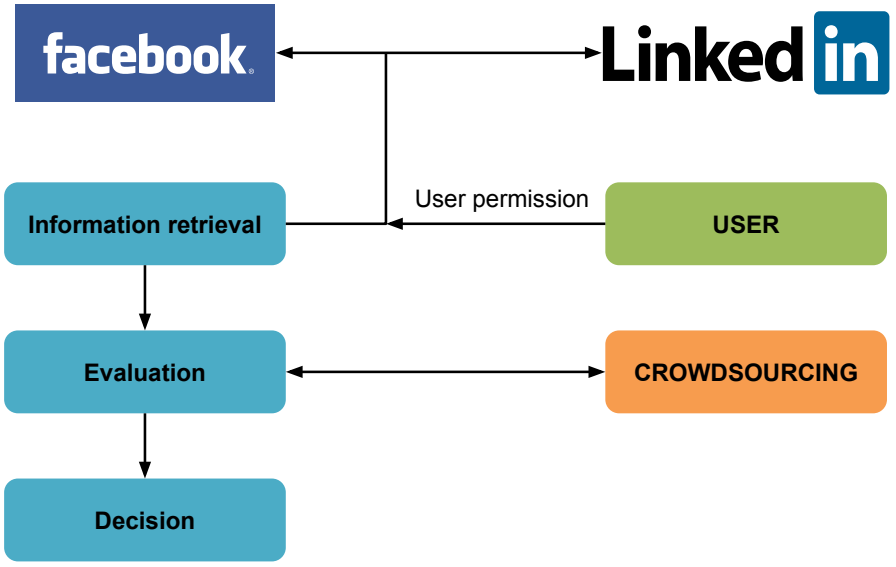

FIGURE 3.1: Risk assessment approach using crowdsourcing and social networks [16].

Their system retrieves information of credit applicants from their Facebook and LinkedIn social media accounts. Information that potentially influences the creditworthiness of a person is extracted from these accounts (e.g., skills, interests, work experience and recommendations). This information is evaluated using a crowdsourcing approach. In this approach, real-world persons are consulted via messages posted automatically on social media asking them how they think that a certain information value (e.g., the skill *Management*) relates to creditworthiness. Facebook was used in the experiment to

post these messages. Persons could respond by 'liking' the message when they think the information contributes to entitling someone to receive a loan. The persons consulted consisted of people interested in banking and financial institutions. A single numerical value is computed using the results of this crowdsourcing evaluation approach by calculating a factor for each category based on the number of 'likes' and then combining these factors using linear regression. A decision is made whether or not the loan should be granted based on this number. Data of a financial institution in Azerbaijan containing credit history of clients who have previously received a loan and have agreed to share their social media data is used to train and test the system. Their system achieved a 92.5% true positive rate and has a false positive and false negative rate of 6.45% and 11.11% respectively.

There are two main differences between the work of Hasanov et al. and out work. Hasanov et al. focus on predicting creditworthiness of persons. However, in this research we focus on predicting creditworthiness of companies. Furthermore, Hasanov et al. first explicitly ask a credit applicant permission to access the data on their social media accounts and hence knows for certain that the information relates to the credit applicant. In this research, we study creditworthiness prediction using data for which it is uncertain whether the information relates to the person or company of interest.

3.3 Credit scoring with social network data

Besides information directly related to a person for whom the credit score is calculated, also the information of social media connections can act as information source. Y. Wei et al. performed an economic study in which they analyze the use of social network data on the accuracy of consumer credit scores [3]. The results of their study show that, under the assumption that people are more likely to connect with people that are similar to them, social network data provides extra information that can improve the accuracy of creditworthiness prediction. Social network connections of a person can be used as additional information for predicting creditworthiness by not purely looking at who those friends are, but by analyzing how those persons are connected. It is reasonable to assume that there has to be some kind of link or similarity between two persons connected on social media (e.g., traits, work or study). This similarity can be used to clarify beliefs about the person and verify its information. Whenever companies start using this social network data for estimating creditworthiness, people might start limiting their number of connections and only connect to more similar people to try to improve their credit score. This can result in network fragmentation in which good financial types only connect with other good financial types, causing bad financial types

only being able to connect with other bad financial types. If people getting rid of bad connections, the system is getting information from less connections, resulting in less information to improve the creditworthiness prediction accuracy. However, when the system can get information and knows that you are trying to game the system by disconnecting bad friends, and it knows that you are a good type, it also knows that your connections are more likely to be also good types. When a group of good types is in your network, the system could have more confidence that you are also a good type. The results suggest that whenever a prediction of creditworthiness can be made using the social media account of a person, the information of connections of this person might also be an indicator of the creditworthiness of the person.

3.4 Probabilistic databases

In this research, we study whether creditworthiness predictions can be made using uncertain data. A lot of research has been done on dealing with uncertainty in databases. These so-called probabilistic databases are database management systems which can store uncertain data [17]. The value of some attributes or the presence of some records within these databases can be uncertain and are only known with some probability [18]. Traditional database management systems store all values as facts. However, there are many sources of uncertainty. Suciu et al. [18] distinguish between two types of uncertainty: tuple-level uncertainty and attribute-level uncertainty. Whether a tuple belongs to a database instance or not, i.e., whether the tuple exists, is unknown in situations of the first type of uncertainty. Each tuple is a random variable which can take two values: true when the tuple is present and false when it is absent. Hence, it is called a maybe tuple. In situations of the second type of uncertainty the value of an attribute A is uncertain. The value of this attribute A is a random variable whose domain is the set of values the attribute might take for that tuple. These forms of uncertainty can come from missing data, non-specific data, vague data, inconsistent data and errors [17, 19, 20]:

- In information extraction systems uncertainty comes from the ambiguity in natural-language.
- In data integration and cleaning uncertainty comes from comparing data sources and identifying matching entities.
- In data collection uncertainty comes from the limited accuracy of measurement equipment.

- In some privacy important applications uncertainty is purposely inserted to hide sensitive attributes.

Examples of probabilistic database systems are MayBMS [21], Trio [22], MystiQ [23] and Orion [24]. The first two only support discrete uncertainty while the last two also support continuous uncertainty. These are all relational probabilistic database systems. However, also research is done to build a probabilistic XML database system [25, 26].

An important concept within the probabilistic databases field is the concept of possible worlds. Because the value of attributes and the existence of tuples are uncertain, the database can be in one of several possible states each of which has a certain probability. These possible database instances are called possible worlds. Table 3.1 shows an example illustrating the possible worlds concept. This example shows a database consisting of a table with the attributes name and gender. For some tuples in this example, there is attribute-level uncertainty for the name attribute, for example, the name of the first person is 'Jan' with a probability of 0.7 or 'Janssen' with a probability of 0.3. Furthermore, in this example the existence of some tuples is uncertain, for example, 'Marieke' might only belong to the table with a probability of 0.6 (and not belong to the table with a probability of 0.4). Given these possible tuples, the database can be in four states. The probability that the database is in state 3.1a equals $0.7 \times 0.6 = 0.42$. The probabilities for the other possible worlds were calculated similarly. This is a relatively simple example with a limited amount of possible worlds, but in more realistic cases the amount of possible worlds can get huge. It is often impossible to return all possible sets of answers. Therefore, the probabilistic database often ranks tuples and aggregates over uncertain values to present the possible query answers to the user [19]. For example, query performance can be speeded up by returning only the k highest ranked tuples (top-k query answering).

TABLE 3.1: Possible worlds example

(A)

Probability: 0.42

Name	**Gender**
Jan	male
Marieke	female

(B)

Probability: 0.28

Name	**Gender**
Jan	male

(C)

Probability: 0.18

Name	**Gender**
Janssen	male
Marieke	female

(D)

Probability: 0.12

Name	**Gender**
Janssen	male

Possible worlds can be defined more formally as follows: let $\mathcal{DB}$ be set of all possible databases. Let D be an instance of a database: $D \in \mathcal{DB}$. Furthermore, let $\widetilde{D}$ be a probabilistic database instance. A probabilistic database is a set of database instances, each associated with a certain probability: $\widetilde{D} \in \mathcal{DB} \times [0..1]$. The sum of the probabilities of all possible states of this uncertain database should equal one: $\sum_{(D,p)} p = 1$. D is

called a possible world of the probabilistic database. The set of all probabilistic databases can then be defined as follows: $\mathcal{PDB} = \{\widetilde{D} \in \mathcal{DB} \times [0..1] | \sum_{(D,p)} p = 1\}$

Probabilistic database systems can support two types of uncertainty: discrete and continuous. A finite number of values can be distinguished with discrete uncertainty while in continuous database systems attributes can take an infinite number of values. Besides this classification of uncertainty, it can also be classified as dependent or independent. Whenever the attribute values are independent, the value of one attribute does not influence the probability of the value of the other attributes. For example, whether the name is Jan or Janssen in the example of 3.1 does not influence the fact whether Marieke is an entry within this table. The data of a probabilistic database is usually stored in normal relational database tables, supplemented with extra attributed and tables that store probabilities, alternatives and dependencies [17].

Queries sent to a probabilistic database are to a large extent the same as standard SQL queries. Most probabilistic databases use an extended SQL version [17]. This extended version of SQL supports some special constructs that allow dealing with uncertainty, for example, for calculating the expected value. However, these constructs are not standardized and differ per probabilistic database system. Instead of only returning the results, a probabilistic database also returns the probabilities of each result. A SQL query on a probabilistic database is computed by fetching and transforming the data; and performing probabilistic inference [19]. One of the advantages of probabilistic databases is that it integrates the probabilistic inference and query computation steps. By doing so, the probabilistic inference can be speeded up by using standard database management techniques such as materialized views and schema information.

Another advantage of probabilistic databases is that it often does not require tuples and attributes to be independent and allow dependencies in the uncertainties to be specified. For example, a tuple can only exist if another tuple exists, or two tuples cannot coexist. Similarly for attribute uncertainty, you can specify, for example, that an attribute has a particular value only if another attribute has a particular value. The advantage for application developers is that they do not have to be concerned with the details probability theory, because the uncertainty is dealt within the database management system. Developers can simply query the database with the slightly different version of SQL.

Queries are evaluated according to an execution plan. Query optimization is concerned with finding more efficient execution plans while preserving the semantics. A query result of a probabilistic database is correct only if both assertions and their probability are correct [27]. A safe execution plan is a query execution plan that returns correct results. Once a plan has proven to be safe it can be executed on any database instance. In contrast to default database systems, tuples within a probabilistic database may

be dependent, and hence cannot be manipulated independently. Therefore, a query plan that is correct for regular database systems might not be correct for probabilistic databases.

In [15], Van Keulen et al. claim that dealing with alternatives (i.e. possible worlds) along with their confidences might yield better results. In this work, we study whether this claim of achieving better results by applying the possible worlds concept of probabilistic databases holds in a real-world case within the financial domain. We study how incorporating multiple candidate results instead of incorporating only the most likely result affect the accuracy of a creditworthiness prediction model.

Chapter 4

Financial Data

Constructing a data mining model for subquestions three and four, that can make creditworthiness predictions based on online information, requires both financial and online data to train and test the model. Since in this research we study the prediction of a company's creditworthiness, the financial data should be related to the creditworthiness of companies. In this chapter, we discuss two possible creditworthiness indicators. We discuss the data provider that is used for constructing a data set consisting of this selected creditworthiness indicator. Furthermore, we explain the setup that was used to collect the data from this data provider. The resulting data set is discussed. This allows us to answer the first subquestion on the available financial data related to creditworthiness to test the performance of online information as predictor for a company's creditworthiness.

4.1 Creditworthiness indicator

There are multiple creditworthiness indicators that could be used for training and testing a model that uses online information for predicting a company's creditworthiness. In this section we discuss bankruptcies and credit ratings as creditworthiness indicators.

4.1.1 Bankruptcy

Bankruptcy is the ultimate proof that a company was not creditworthy, because a company that went bankrupt cannot repay its loan. Figure 4.1 shows how a test setup might look like when using bankruptcy as creditworthiness indicator for testing the performance of online information as predictor of creditworthiness. Although this setup also includes a rating, it is not used for determining the accuracy of the prediction. In this

test setup the accuracy of the original rating model is determined using the bankruptcy information of companies. This accuracy is used as a baseline. A new rating model that incorporates online information into the original rating model can be constructed. Some basic information about the company, that is available through a data provider, can be used for searching more elaborate information about them in an online information source. This information is transformed into an indicator score, which can be added to the original model. The accuracy of this new rating model can again be determined using bankruptcy information. The accuracy of both the model excluding and including the online information can be compared to determine whether the online information improves the creditworthiness prediction.

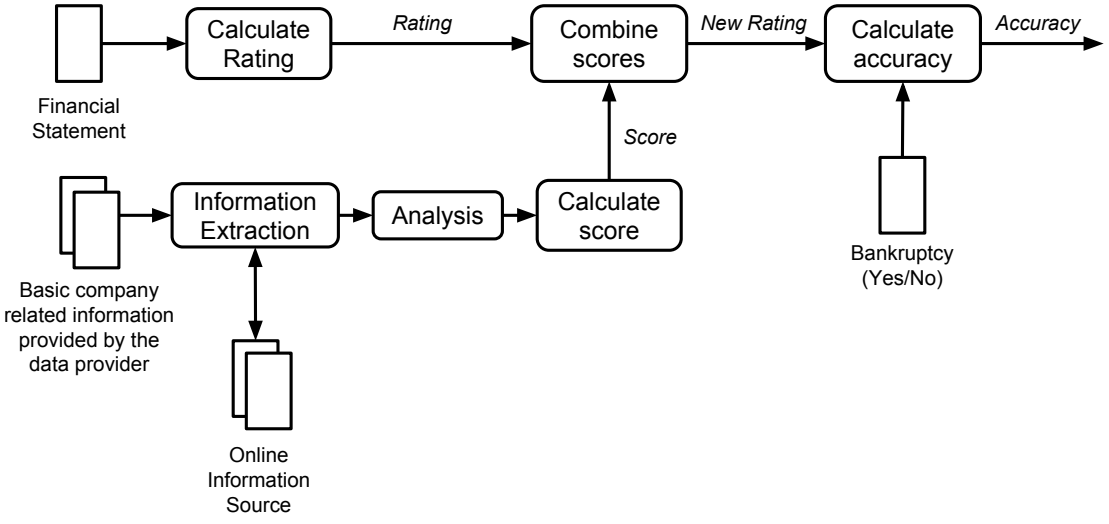

FIGURE 4.1: Test setup when using bankruptcy as creditworthiness indicator.

Several problems were discovered when attempting to create a data set in which bankruptcy is used as creditworthiness indicator. The results of a manual verification of the availability of online information about several companies that went bankrupt and persons that are related to these companies suggest that:

- In general, companies that went bankrupt do not seem to be very active on the Internet. Many of these companies do not have a website and could not be found on several popular online platforms such as Twitter, LinkedIn and Facebook. In addition, persons related to these companies seem to be harder to find on the Internet.
- Even if the companies were active on the Internet, their data often is limited, outdated or no longer available.

The presence of companies on the Internet might by itself be predictors of creditworthiness. However, when no account can be found for a company and hence no extra information can be extracted from it, the presence on the Internet would be the only indicator that could be used for making predictions. The performance of other online information as predictor of creditworthiness cannot be tested in these cases. In addition, often the amount of information available for these defaulted companies or persons that are related to it is too limited to determine irrefutably whether the account belongs to the company / person and hence whether it present on the Internet. Since in these cases it is uncertain whether the company / person had an account, it cannot be used for training and testing the model. Because of these issues it was decided not to use bankruptcy as creditworthiness indicator.

4.1.2 Rating

A rating expresses the probability of default and indicates how likely it is that someone is able to repay its loan within a certain period. As discussed in Section 2.2 a rating can be calculated based on the figures of the financial statement and profit & loss account. Using ratings as creditworthiness indicator does not necessarily suffer from the same problems as when using bankruptcy as creditworthiness indicator. First, it might be easier to find a set of companies that are relatively less creditworthy, but still can be found on the Internet. Most bankruptcies occur among smaller companies because they are more volatile and more often operate in small markets [28]. These smaller companies are relatively harder to find on the Internet. The amount of companies that went bankrupt is a subset of the amount of companies that would have had a bad rating. Hence, there might be more relatively larger companies with a bad rating for which information can be found compared to the amount of companies that went bankrupt for which information can be found. Second, most companies for which a rating can be calculated are still active. Therefore, it is more likely to be able to find up to date information for these companies.

Using ratings as a creditworthiness indicator requires a slightly different test setup compared to the one discussed in Section 4.1.1. In general, testing cannot be done with data that is also used to construct a model. Hence, when the calculated score would be combined with the rating as in Figure 4.1, the rating cannot also be used for testing the accuracy of the score because it partly already consists of this rating. However, another possibility is to use the rating as a reference and try to approximate this rating by calculating a score over the online information. Figure 4.2 shows how a test setup might look like when using the rating for testing the performance of online information as predictor of a company's creditworthiness.

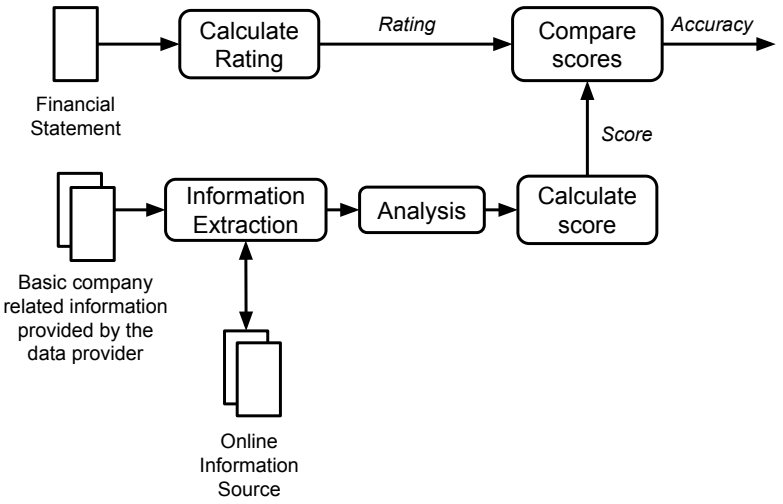

FIGURE 4.2: Test setup when using ratings as creditworthiness indicator.

4.2 Data provider

A data provider needed to be found that could provide ratings or the financial data necessary to calculate ratings, such that ratings can be used for training and testing our prediction model. Interviews with several employees of Topicus Finance revealed an external data provider that could supply the information necessary to create a data set. This data provider offers company information on millions of listed and private industrial companies around the globe. It cannot provide company ratings off the shelf. However, Topicus Finance has access to a rating formula that uses the figures of the financial statement of a company to calculate a credit rating. Since the data provider can provide key financial data for the companies, such as balance sheet and profit & loss account figures, it is possible to compute a rating for them. In addition, if listed, the data provider also offers other relevant non-financial information such as the company's contact information, the legal status (e.g., active, default on payment or bankrupt), the sector it belongs to; and names of current directors and managers. This additional data allows online information related to these companies to be searched.

4.3 Collecting data

To construct a data set consisting of ratings that could be used for training the prediction model and testing its performance, it was necessary to build a system that could import and extract the data from the data provider. As explained in the previous section, the data provider does not provide ratings itself. However, ratings could be calculated over the data of this data provider using the rating model available within Topicus Finance. We will now discuss the setup of this system.

4.3.1 Setup

Figure 4.3 illustrates the setup that was used to create a data set using the data provided by the data provider. Non-shaded components were already in place within Topicus Finance while the shaded components were specially created for this system. The different components within this setup are discussed together with the flow of information.

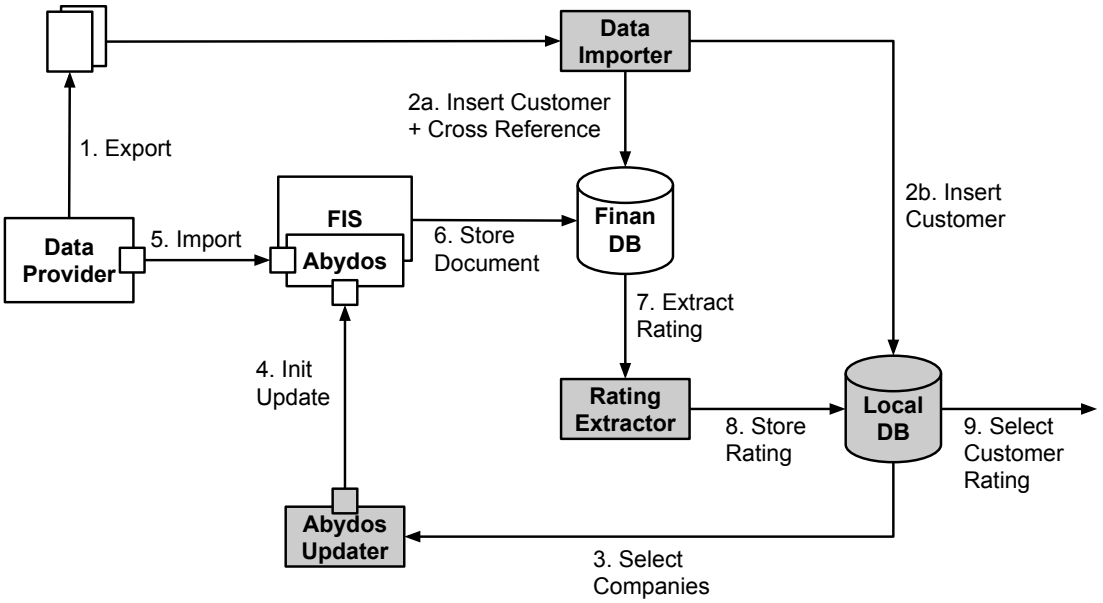

FIGURE 4.3: Setup for creating data set.

- **Data Provider**. As discussed before, the data provider provides financial information together with some basic information about the companies. The financial information is accessible through a web service while the basic information can be exported (1) though the data provider's web interface. This basic information consists (if available for the company) of:
  - Company name
  - Company managers / directors / contacts
  - Address: street, zip code, city, country
  - Website
  - Email
  - Statistical Classification of Economic Activities in the European Community (NACE)
  - Chamber of commerce registration number
  - Company status (including the date of this status)

- **Data Importer**. The Data Importer is designed to read files that are exported by the data provider and storing the exported information in the Finan and Local DB. A customer entry and cross reference are created and stored in the Finan DB (2a). The customer details are stored in the Local DB (2b).

- **Finan Database**. The three elements stored within the Finan DB for this setup are customers, cross references and documents. Customers consist of a name and reference identifier which is used to access and store information of the customer within the Finan system. Cross references belong to a certain customer and contain the identifier used by the external data provider. This identifier is used to retrieve financial data of a specific customer. Documents are stored in XML format and also belong to a certain customer.

- **Local Database**. The Local DB stores both customer details and ratings that are extracted from the imported XML documents.

- **Abydos Updater**. The Abydos Updater initiates the retrieval of financial documents. It selects customers that have been added to the Local DB (3) and initiates updates by calling the *InitUpdate* method of the Abydos web service (4).

- **Abydos Web Service**. The Abydos web service can be used to initiate updating financial data stored within the Finan DB. Abydos also has a web interface which allows running update batches. However, updates initiated through this web interface will only update documents for which the data provider's data has changed in the last year. In contrast, updates initiated through the web service are not only limited to documents updated in the last year and will update all documents for which there is a newer document available.

- **Financial Import Service (FIS)**. The FIS is a component used within Topicus Finance to import financial data. The Abydos web service uses this service for importing the updated financial data. The FIS first retrieves the financial XML data from an external source (5). Then it transforms this data using a defined mapping into a format that is understood by the Finan model. In this case, it transforms the data into International Financial Reporting Standards (IFRS) format which is an accountancy standard. The Finan model was modified for this setup such that ratings are added to the XML documents stored in the database. Finally, it stores this document into the Finan DB (6).

- **Rating Extractor**. The Rating Extractor extracts ratings from XML documents stored in the Finan DB (7). This is done using SQL/XML queries. SQL/XML is an extension to the SQL language which allows to store and query XML data in a relational database [29]. Figure 4.4 shows the relevant nodes of the XML tree of

these financial documents. The relevant nodes are: column definitions, timelines and financial nodes. These are contained within the *columnDefinitions*, *timeLines* and *ROOT* nodes respectively. Column definitions define a period to which the financial figures relate. The XML document has a separate timeline for every information source. In most cases, there is only a single timeline since there is only one source of information. Each timeline consists of a number of columns for each year for which there is financial data available. The financial nodes (in this case *krPD* and *krUraRatingKlasse*) consist of column nodes, which relate to the columns defined within the *columnDefinitions* and contain the specific financial value.

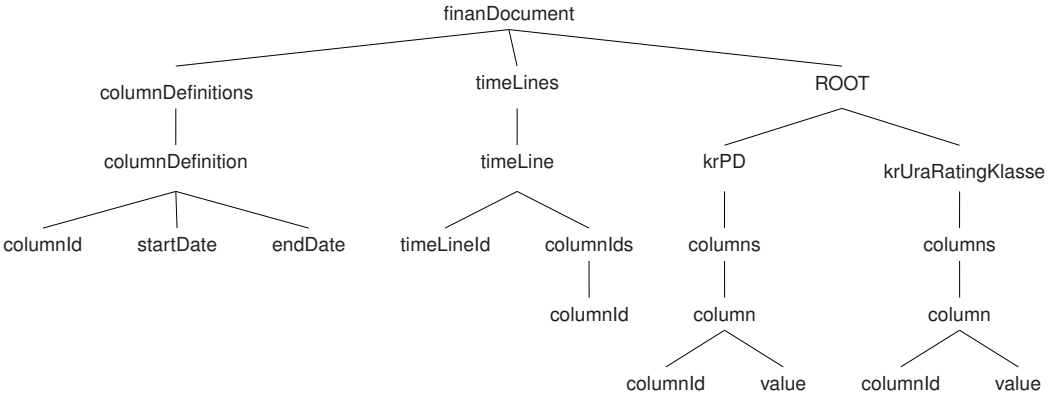

FIGURE 4.4: XML tree of the financial documents.

4.3.2 Data filtering

The data provider offers company information for a large variety of companies. However, the data provided is not always sufficient for a rating to be calculated. Furthermore, not all companies are within the scope of this research. Therefore, some selection filters were applied before importing the data to increase the quality of the data set. These filters were chosen with care to not unnecessarily limit ourselves, but whenever it turns out that some of these filters are too strict they can easily be loosened. The data filters applied initially when selecting data from the data provider are:

- **Incomplete data**. The financial data of many companies provided by the data provider is incomplete and therefore insufficient for a rating to be calculated. Hence, filters were added to select only companies for which the balance sheet contains at least the amount of total assets and shareholders' funds & liabilities; and for which the profit & loss account contains at least an operating revenue value.

- **Country**. Importing and extracting the necessary data to create a data set is time consuming. Because of time constraints, the first step was to create a data set consisting of only Dutch companies. This country is from the point of view of Topicus Finance most interesting.
- **Latest year of account**. Only companies for which the financial figures are at least available for one of the five most recent years (2010-2014) are selected. This filter is essential, because it excludes companies for which only data is available when social media did not yet exist or the usage was still limited. In addition, it limits the number of ratings which are not related to information that can be found today.
- **Financial companies**. The mappings created by Topicus Finance are not intended for financial companies. These types of companies have an accounting template which deviates from the industrial accounting template. Therefore, these companies are excluded (NACE Rev. 2 nr. 64-66 and 69).
- **Other irrelevant sectors**. Public administration and defense; and extraterritorial organizations and bodies are also excluded, because these companies do not follow the credit application process.
- **Exclude holdings**. Because holdings do not produce any goods or services, they are excluded. In addition to the financial holding which already is excluded, this also concerns non-financial holdings (NACE Rev. 2 nr. 70). Companies with the word 'Holding' in their name were later removed, because it turned out that the sector filter did not filter out all holdings.
- **Legal form**. Companies with a sole proprietorship, private limited liability and general partnership legal form are selected. Public limited liability companies are excluded from the data set. Often, much information is available for these types of companies to make already a good prediction of their creditworthiness. The potential added value of online information is larger for smaller companies for which there is relatively less other information available.
- **Turnover**. A limit of 50 million euro was set on the turnover for the same reason as why the legal form filter is added. This amount was chosen in consultation with the Topicus Finance economist.

After running the program, the information of 10907 companies was imported. In total, 26703 ratings related to the period 2010-2014 were calculated and extracted for these companies. However, it turned out that this initial data set still had some deficiencies. Therefore, some additional filters were applied after importing the data to further increase the quality of the data set:

- **Unbalanced balance sheets**. When studying the imported financial data, we found that for many ratings the corresponding balance sheet is unbalanced after importing the data. An inspection of these balance sheets revealed that the mapping would not map certain figures to the Finan system when it lacks a level of detail, i.e., no subfigures are defined for these main category figures. We found that this mapping could be improved by additionally mapping the main category of these subfigures. Ratings calculated over unbalanced balance sheets are likely to be incorrect. Therefore, an additional filter was applied to the already collected data that checks whether the balance sheet is balanced. When this is not the case, the rating is left out of the data set.

- **Director / manager / contact information available**. Because we are specifically interested in whether online information about a person related to the company can be used as a predictor of creditworthiness, it is necessary that some basic information about these persons is known. Therefore, only companies for which this information was available were selected.

- **Extreme ratings**. Figure 4.5 shows the distribution of 7703 ratings for which the balance sheet is balanced and for which information about the company directors / managers / contacts is known. The ratings seem to be normally distributed. However, as can be seen in the figure, the AAA and especially the D ratings seem to be outliers to this distribution. Despite being interested in the difference in ratings, the Topicus Finance economist advised to filter out these extreme ratings. A common reason for companies to obtain an AAA rating is that the company has a high amount of equity. According to the economist, many of the smaller companies with a high amount of equity are investment companies. Although on paper these companies might not be classified as holdings, in practice they do not produce any goods or services and therefore are out of scope for this research. A company receiving a D rating is an indication that it is quite certain that the company will go bankrupt within a year. If the rating concerns a period more than one year ago and the company is still active, the rating has a high likelihood of being incorrect. For example, 4% of the companies had a probability of default of 100% in one year and still existed in a later year. This means that at least 20% of the D ratings in this set were incorrect. If these ratings are kept included, the system would be trained and tested on erroneous data and as a consequence will produce inaccurate results.

- **Most recent ratings vs. all available ratings**. When the company ratings of all available years are used to determine how a change in information influences the rating, it is important that we can identify which information was available

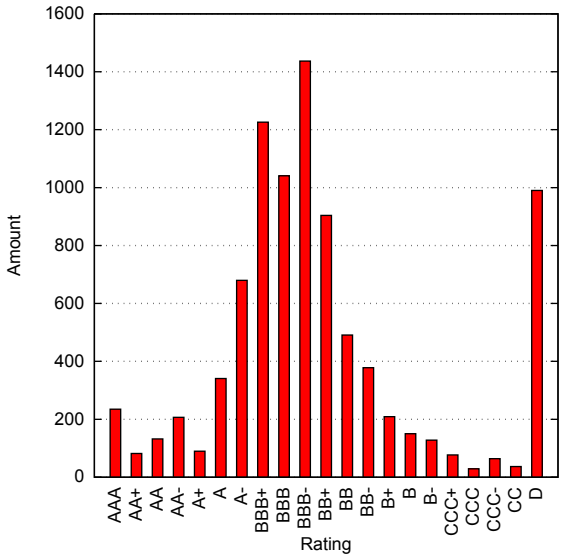

FIGURE 4.5: Rating distribution including multiple ratings per company.

at the time of the rating and which was placed online later. Often, the creation or modification date of information on the Internet is unknown. Whenever only the most recent ratings are used, we limit the number of cases for which the information that is now available was unavailable at the time of the rating. Hence, it was decided to only use most recent ratings.

4.4 Result

After filtering the data set, the ratings of 3579 distinct companies remained. Figure 4.6 shows the distribution of most recent ratings available per company for which the corresponding balance sheet is balanced and for which information about the company directors / managers / contacts is known. For training and testing our prediction model, the data set should consist both of companies that are creditworthy and those that are not such that the difference in information can be related to the difference in creditworthiness. The figure shows that the ratings of this data set are indeed well-distributed.

In general, having a bigger data set is better for data mining. However, the costs in time, money and effort should we weighted against the benefit of extending the data set. The models that will be constructed are trained on data for which uncertainty is not modeled, i.e., the information is actually related to the company belonging to a rating. To be certain that harvested information is related to a company, it is necessary to do a manual search for the correct pages / accounts within online information sources. In our case, the costs are the time and effort necessary for performing this manual search.

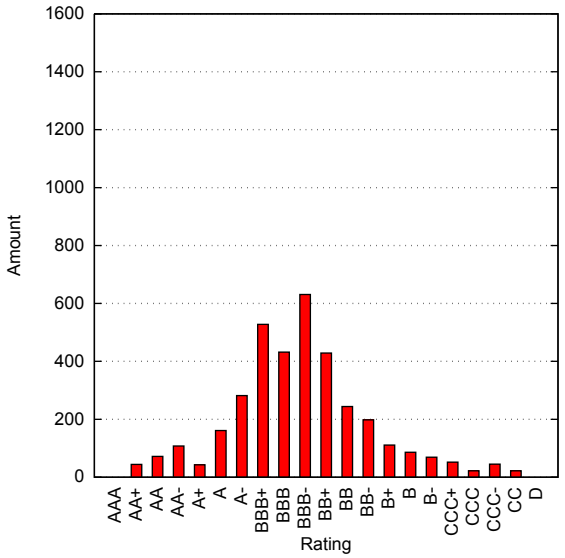

FIGURE 4.6: Rating distribution after refining the data set.

Because of time constraints, we might not be able to search in all considered online information sources for the pages / accounts of all 3579 companies. No specific criterion was found for the minimum size of a data set in data mining, because it strongly depends on the experiment that is performed, and the data and classifier that is used. However, related work [16] suggests that 400 is a reasonable amount. And although in this research we do not take a statistical approach, a sample size of 385 seems to be required for an unknown population and confidence level of 95% [30]. The final data set amply met this criterion. This suggests that this data set can be used for the experiments.

The figure shows that fewer samples are available for the outer rating classes. Whenever only a subset of these companies is manually searched, it might turn out that there are too few samples for some rating classes to train the classifiers. However, given the large number of rating classes, we can reduce the number of classes, such that more samples per class are available whenever the amount of samples per class turns out to be a problem.

4.5 Conclusion

In this chapter, we discussed which financial data is available for training and testing a model that uses online information for predicting a company's creditworthiness. Bankruptcies and ratings were discussed as possible creditworthiness indicators for training and testing a creditworthiness prediction model. Several problems of using bankruptcy as creditworthiness indicator were discussed. It was explained that using ratings as creditworthiness indicator suffer not or to a less extend from these problems.

A data provider was discussed which could provide the financial data necessary for using ratings as creditworthiness indicator. The setup that was used to import balance sheet and profit & loss account figures from the external data provider for a set of Dutch companies was explained. Ratings were calculated over these financial figures using a rating formula to which Topicus Finance has access. After applying some selection filters to improve the quality of the data set, a set consisting of well-distributed ratings of 3579 different companies remained and is available for testing the prediction performance. Related work [16], in which a data set of 400 entries is used, suggests that this is a sufficient amount for training and testing the prediction models in the experiments of the third and fourth research questions.

Chapter 5

Online Data

In addition to the financial data discussed in the previous chapter, online data is required for testing the performance of online information as predictor of a company's creditworthiness. In this chapter, we study which online data could be used for making these predictions. First, the results of a survey, which was held to identify online information sources which according to employees of Topicus Finance can be used for making creditworthiness predictions, are discussed. Next, literature related to information sources for creditworthiness predictions is discussed. To be able to use an information source for training and testing a data mining model, it is required that for a significant amount of companies from our financial data set information can be found in the information source. Hence, a test was performed, in which the presence of companies in several information sources is tested. The most promising information sources and corresponding information are determined based on the results of these availability tests. Last, we explain how the online data necessary for training and testing the model is collected. This allows us to answer the second subquestion on the available online information that can be used as a predictor for a company's creditworthiness.

5.1 Survey

A survey was held within Topicus Finance in the exploratory phase of this research to identify possible interesting online information sources and information subjects. This survey mainly focuses on social media as an information source, because of the results of Heijnen [4] which show that social media usage is considerable for businesses of various industries. However, also some questions were asked to identify additional interesting information sources. The survey was held among employees of Topicus Finance. In total 17 people filled in the survey. The survey can be found in Appendix D. Additional

interesting information sources were identified using an open question. The results of all multiple choice and open questions of this survey can be found in Table E.1 and E.2 in Appendix E respectively. The main results are discussed below.

Main Results

The results of this survey confirm that social media websites, such as Facebook, LinkedIn and Twitter, might be interesting as an information source. The information subjects about which the survey participants would search information in the role of a lender were prioritized as:

1. The company in question.
2. Company owner(s).
3. Employees.
4. Other companies within the same sector.
5. Family of the company owner(s).
6. Friends of the company owner(s).

Additional information subjects that were identified as possibly interesting from the perspective of a lender are:

- Suppliers & customers
- Other financiers
- Country
- Past owners

On Facebook, the participants of the survey seem to be primarily interested in message content. Information they would search for within these messages is: company performance; remarkable information; treatment of employees; complaints and compliments; and whether the messages can be classified as spam or are a more serious attempt of the entrepreneur to build up a network. According to the participants, LinkedIn mainly seems to be interesting because of the skills, experience and education of persons. On Twitter, the survey participants seem to be more interested in the amount of followers and amount of retweets as a measure of popularity and broadcast radius. There also

seems to be some interest in messages on Twitter. They are mainly interested in the sentiment within these messages (amount of complaints and compliments). Other online information in which the participants are interested in consists of:

- Company website.
- Whether the company or entrepreneur actually is present on social media.
- The behavior of the entrepreneur on Internet (moral/communication expressions/character).
- Regulations. From the survey it did not become clear whether people were interested in certain specific regulations, or more generally in the regulatory burden.
- Sector outlooks published by banks or statistics bureaus.
- The living area of the entrepreneur (e.g., does the entrepreneur life in an underprivileged neighborhood?).
- How well the company is reviewed (e.g., on Google Reviews).
- The amount of publicity on news websites and the nature of this publicity.
- Number of outstanding job vacancies of the company.
- Metadata (e.g., how often a website is updated).

5.2 Literature

In this section, literature is discussed that was studied to determine which information is used in similar problems, in which sources information related to companies can be found and what information could be found.

5.2.1 The added value of external data sources in the credit application process

One of the products that have been developed by Topicus Finance is Finan Online. This is a credit application tool in which an entrepreneur or accountant can prepare a credit application in eight steps. Within these eight steps, information about the company, its environment and finances is entered. M. Brinkhuis improved one of these steps in which the entrepreneur or accountant needs to fill in information about the company's environment [31]. Brinkhuis identified several sources that can help the user to fill in

this step and included some of these sources in the Finan Online application to better guide the user in filling in the data. The information that he included was:

- **Market information**. This includes an overview of sector news items, sector statistics (e.g., the number of bankruptcies and number of job vacancies within the sector) and market size statistics (e.g., consumer income).
- **Competition information**. This includes a selection of competitors and a manually written note on why they are competitors and how the company performs in comparison to these competitors.
- **Environment information**. This includes announcements published by the local government and statistics about the region (e.g., the number of bankruptcies and number of vacancies within the region)
- **Company information**. This includes information on company finance compared to other companies within the same sector and the personality of the entrepreneur.

A major drawback of this solution is that it is still necessary for users to manually interpret each of these data sources themselves and need to explain in their own words how it affects their company and therefore also indirectly their likelihood of obtaining a credit.

5.2.2 Online credit score calculating companies

The results of Hasanov [16] discussed in Section 3.2 of Chapter 3 suggest that social media can be used for making creditworthiness predictions. This is supported by the existence of companies that use social media data for creditworthiness predictions of persons. As mentioned in Section 1.2, in recent years, several companies have emerged that use information of the Internet for calculating credit scores. Websites used by these companies are mostly social media and include Facebook, Twitter, LinkedIn, Gmail and Yahoo [6]. The precise working of the algorithms used by these companies remains unknown, because it is part of their business model. However, sometimes some information was revealed by the owners about the information sources that are used:

- Lenddo Inc.[1] is such a company that uses a person's information on social networks to compute a credit score before granting a loan. It uses profile information such as education and career data, the amount of followers they have, the connections they have, and information about these connections [2].

[1] https://www.lenddo.com/

- Similarly, Neo Finance Inc.[2] has been reported to use the information of LinkedIn profiles to "determine how long users have held jobs, the number and quality of connections in their industry and geography and the seniority of their connections" [2].

- Friendly Score Ltd.[3] uses information of both Facebook and LinkedIn to calculate a credit score that can be used as an extra indicator by credit lenders. They trained their system to identify features that correlate to populations of good borrowers and bad borrowers. According to the founder of Friendly Score, one of such features is tagging your family members among your friends circle [6]. The similarity of a person to those people who pay back their loans or not is calculated to obtain a credit score.

- Affirm Inc.[4] also asks their users to connect their Gmail or Facebook accounts, which they use to verify the identity of the user. After the identity has been verified, their system scans a large set of available data associated with the identity to calculate a credit score. Social information across social media is combined with marketing databases and credit history information [32]. The social information used within the credit score algorithm again includes a person's location and the amount of connections [2].

- The aforementioned companies focus on consumers and do not relate the social information to good entrepreneurship. However, also business oriented credit granting companies exist. Companies like Kabbage Inc.[5] focus on corporate loans and also take more financial related website, such as Amazon and PayPal into consideration into their credit scoring algorithm.

5.2.3 Social media as information source

As discussed before, social media is an important potential information source that could be used for creating creditworthiness indicators. According to research of the Dutch central statistics bureau, over 50% of Dutch companies and more than 80% of the Dutch people is active on social media[6]. In this subsection literature related to social media as an information source is discussed.

[2] https://neoverify.com/
[3] http://friendlyscore.com/
[4] https://www.affirm.com/
[5] https://www.kabbage.com/how-it-works/
[6] http://www.cbs.nl/nl-NL/menu/themas/bedrijven/publicaties/artikelen/archief/2015/gebruik-sociale-netwerken-sterk-toegenomen.htm

Social media metrics

Social media can be used to study a large variety of metrics. J. Heijnen gives an overview of social media metrics and intelligence that existing social media monitoring tools analyze [4]:

- Volume of posts: is computed by counting the number of messages containing a certain name, such as a person's or firm's (product) name. This number indicates to what extent a company is subject of discussion on social media.
- Engagement: is the level of involvement of people in the brand and is often measured by counting the amount of likes, followers, shares or retweets. However, this can also easily be manipulated.
- Sentiment: is analyzed by determining the attitude expressed by users in social media messages. Often, messages are either classified as positive, neutral or negative. Classification is done by scanning for certain words or phrases that are associated with a positive or negative attitude.
- Geography: is analyzed by determining the geographical location or region at which messages related to some person or firm are posted.
- Topic and theme detection: is analyzed by determining the topics discussed by people in social media messages related to a person or firm.
- Influencer ranking: is analyzed by determining the amount of followers an author of a social media message has.
- Channel distribution: is studying to what degree a person or firm is subject of discussion on the different social media platforms.

Klout Inc.[7] is an example of a company that calculates influence scores of social media users. It does this by using information of social network sites such as Facebook, Twitter, LinkedIn; and websites such as Bing and Wikipedia. It values the number of reactions a user generates compared to the amount of content it shares; how selective the people who interact with user's content are; and the amount of unique individuals that engage in the user's social activity[8].

[7] https://klout.com/
[8] https://klout.com/corp/score

Activity of companies on social media per industry

Not all companies are equally active on social media. Heijnen manually studied public social media messages related to companies from websites such as Twitter, Facebook and (Wordpress) blogs [4]. He found that the average daily mentions differs strongly from firm to firm and that it therefore is not possible to perform social media analysis for all companies, since not enough data is generated. The amount of data generated also seems to differ per industry. Figure 5.1 shows an overview of his results.

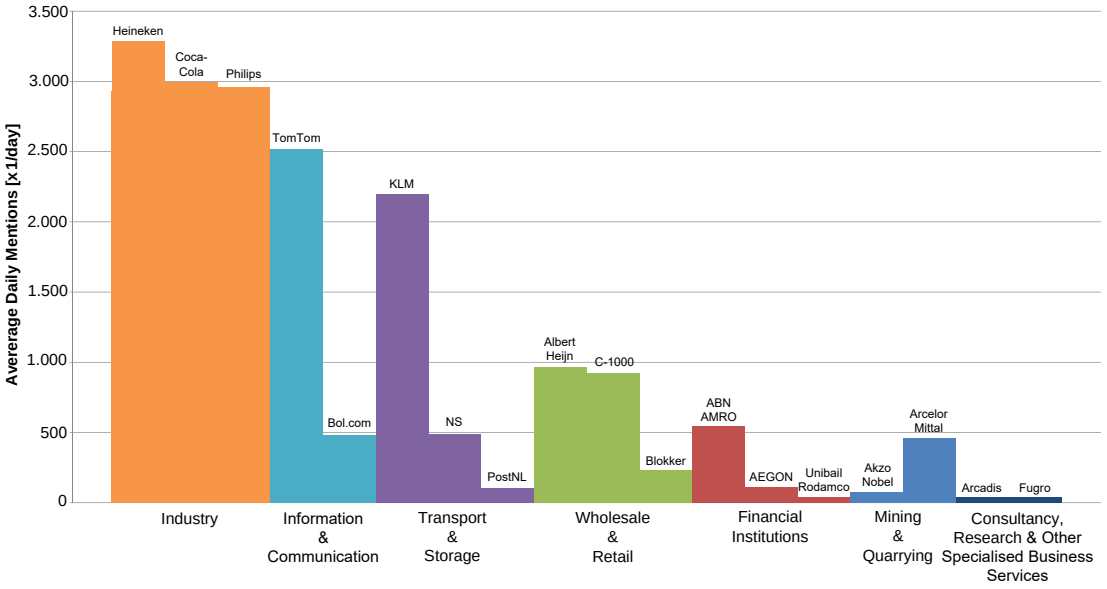

FIGURE 5.1: Average daily mentions of firms [4].

The average daily mentions is clustered per type of industry:

- Industry: producers of food, beverages, chemical products, pharmaceutical raw materials, metal, electric products, etc.
- Information and communication: publishers and/or distributors of books, software, films, music and television shows, etc.
- Transport and storage: transport persons or products.
- Wholesale and retail: companies trading in food, machinery, consumer products, etc.
- Financial institutions: banks, investment institutions, insurance companies, pension companies, etc.
- Mining and quarrying: extractors of oil, gas and/or minerals, etc.

- Consultancy, research and other specialized business services: law firms, accountancy firms, engineering firms, architects, etc.

The results seem to suggest that industrial companies are mentioned more often on social media than for example consultancy companies.

Subjects discussed in messages related to companies on social media per industry

The subjects discussed in messages related to companies also differs [4, 33]. Heijnen et al. manually analyzed social media messages and related them to certain performance indicator categories. These categories include short-term financial results (financial performance discussions and stock related discussions), customer relations (such as questioning, complaining, thanking, explaining and informing), community (such as promotion, news and public image) and other (such as employee relations; operational performance; product and service quality; alliances; supplier relations; environment performance; and product and service innovation). The *other* category consists of subjects that Heijnen et al. found to be under-represented in social media messages. 41% of the business related social media messages he studied were about the perception of stakeholders (e.g., customers) about the company. 18% of the business related social media messages did not contain any useful information. Furthermore, 11% of the social media messages contain financial information (financial performance and stock related discussions). Figure 5.2 shows that the distribution of performance indicator categories of social media messages differ per industry category.

The results of Heijnen et al. show that business-to-business (B2B) companies are less likely to find any useful and new information in social media messages related to the company. In contrast, business-to-company (B2C) firms are more often subject of discussion and these discussions contain new information. Heijnen et al. suggest for future work to automate the classification of messages into a certain category using their manually annotated data set and applying machine learning techniques. However, training probably needs to be done per industry category for accurate results.

Extracting human characteristics from messages on social media

Techniques exist to extract human characteristics from messages on social media. Schwartz et al. [34] describe a differential language analysis (DLA) approach to perform social media analysis. Their technique uses messages posted by persons on social media to find words, phrases and topics that correlate to certain human characteristics such as

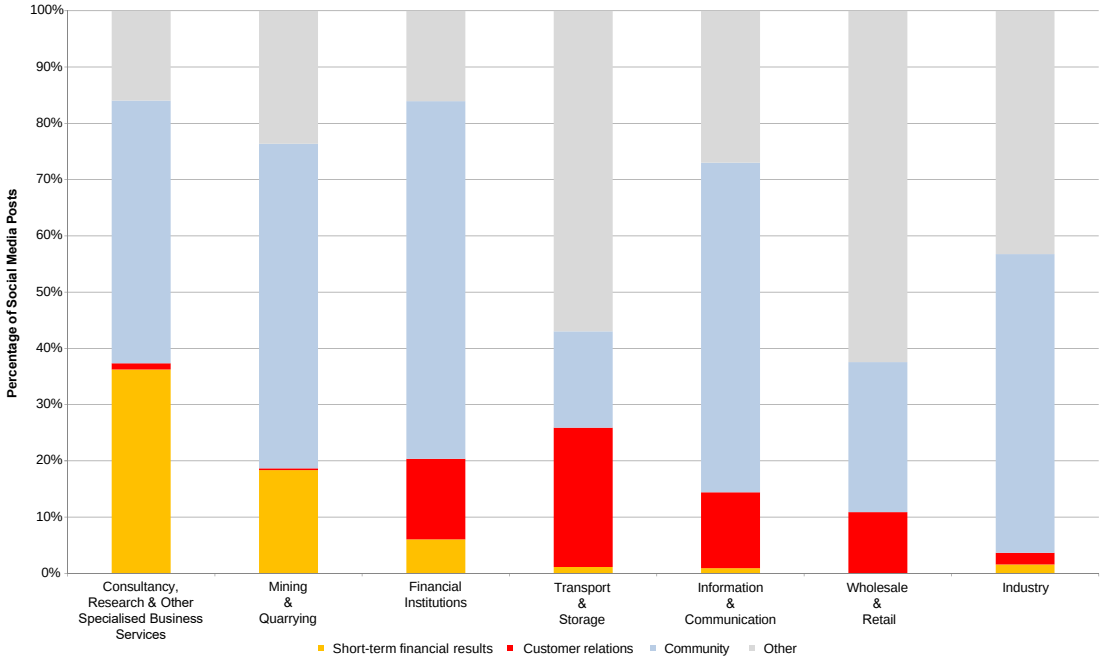

FIGURE 5.2: Social media posts related to performance indicator categories per industry [4, 33].

gender, age and personality traits. Their technique is an open-vocabulary approach in which the correlation analysis is data-driven and not is performed using predefined word lists. Figure 5.3 shows the infrastructure of their system. This system consists of three components:

1. Linguistic Feature Extraction: words and phrases are extracted from social media messages using an emoticon-aware tokenizer. Linguistic features extracted from this text includes: the tense of the text, the perspective in which the text is written and the use of swear words. Furthermore, topics are extracted using a Latent Dirichlet Allocation (LDA) model which assumes that a document is a set of topics each having a certain probability of generating a particular word [35].

2. Correlation analysis: the correlation between each linguistic feature and each human characteristic is determined using ordinary least square linear regression. This is a popular technique in statistics, which attempts to fit a linear function to observed data, i.e., the sum of the squared deviations from the actual data-points to this function is minimized [36].

3. Visualization: the correlation between words and certain human characteristics is visualized using word clouds in which the size of words indicates the strength of the correlation, i.e., larger words indicate a stronger correlation.

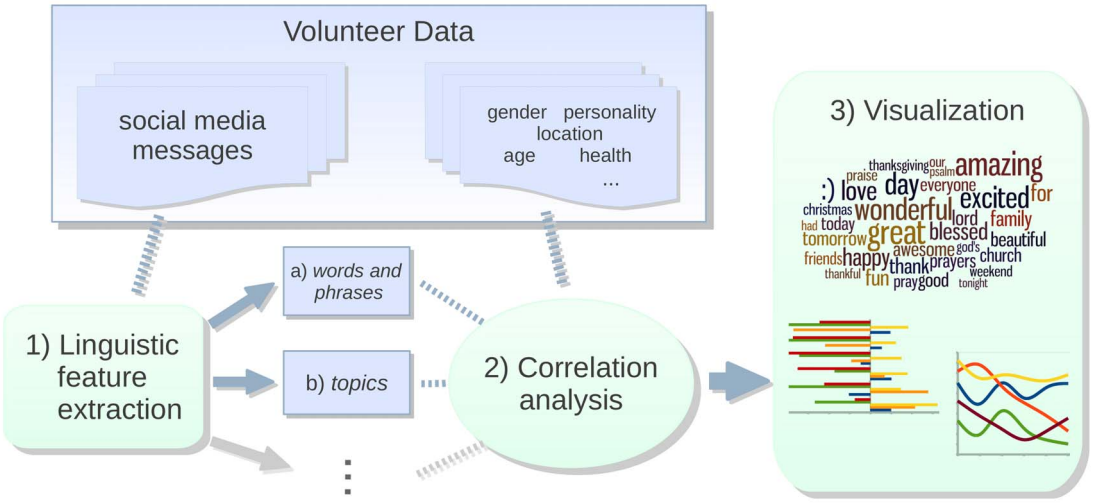

FIGURE 5.3: Differential language analysis infrastructure [34].

Research has been done on the accuracy these characteristics extracted from messages on social media as a creditworthiness indicator. R. Gerrits used the results of Schwartzt et al. to perform an experiment in which the personalities of defaulted and non-defaulted entrepreneurs are extracted from Twitter messages [34]. He studied whether the difference between defaulted and non-defaulted entrepreneurs can be explained and used for a credit rating, based on data from social media [37]. First, he performed a literature study on which kind of personality traits are necessary for a good entrepreneur. He finds that a good entrepreneur should have a high level of openness to experience, conscientiousness and extraversion; but a low level of agreeableness and neuroticism. The personality scores extracted from the social media messages using the DLA approach are compared to these ideal entrepreneur personality scores. Gerrits finds that it is possible to identify, based on social media messages, whether or not a person has personality characteristics of a good entrepreneur which is less likely to default. He confirms that entrepreneurs that go into default have a lower score on openness, conscientiousness and extraversion. However, these entrepreneurs do not score higher on agreeableness and neuroticism. These results suggest that messages of social media can be used for making creditworthiness predictions.

Reputation management

Related to the use of online information (in particular from social media) is online reputation management. For companies, the reputation affects the sales of products and services; and the ability to attract investors and hire new employees [38]. Hence, reputation is important for a company. Reputation is about how others perceive a subject. Doorley et al. [39] define company reputation as: $reputation = Performance +$

Behavior + Communication. Social media focuses mainly on the behavior and communication aspects of reputation. Companies proactively try to improve their reputation through online reputation management. This includes monitoring, analyzing and influencing the reputation of a company, brand or person on the Internet to create and defend a positive public perception [40–43]. Although the intentions of reputation management in many cases might be legitimate (e.g. improving the company's service), it can also be misused. Some commonly used reputation management techniques are:

- Responding to criticism in a timely manner before it becomes widespread [44]
- Search engine optimization and publishing positive content to push negative content downward in the search results[45]
- Sending take-down notices to remove unwanted negative content[46]
- Sponsor others to write positive content about the company or negative content about competitors while masking the sponsorship (i.e. astroturfing)[46]. This is forbidden by the European Unfair Commercial Practices Directive [47]

It is out of scope of this research to validate the sincerity of content. However, a lot of research is being done on detecting fake accounts[48] and (opinion) spamming[49–52]. In addition, websites such as Facebook use and constantly improve their techniques for detecting content manipulation [53].

5.3 Information uncertainty

This research is about the impact on accuracy of accounting for the uncertainty of online data. Therefore, online information sources in which there is some level of uncertainty are of main interest. Two distinct types of uncertainty are identified:

1. Reference uncertainty. Is the information actually related to the company, or does it for example belong to a company with a similar name? Similarly, when searching for information on persons, does this information actually belong to that person? The level of reference uncertainty can be determined using similarity functions or a Named Entity Disambiguation system (See Appendix A and B). In Section 7.2 of Chapter 7, we explain how a search system could be used to determine the level of reference uncertainty.
2. Data uncertainty. Is the information itself uncertain? Are the information values exact or do they have a certain known confidence interval?

5.4 Information subjects

Two approaches can be taken to gather information about a certain information subject. The most obvious approach is to directly search data about the subject. Although this information might be most relevant, it can also be biased since it is easy for an information subject to manipulate the information about him or herself (i.e., reputation management). In general, it is more difficult to manipulate information about subjects that is produced by others. Three categories of online information subjects that might be relevant for a company's creditworthiness prediction are identified:

1. Company. The most obvious information subject is the company for which the creditworthiness needs to be estimated. However, it might also be interesting to search for information on (local) competitors and suppliers.

2. Person. The most obvious persons to search for to estimate the company's creditworthiness are the directors and managers. Information about these subjects might give an impression of how well the company is managed. For example, do these persons have good qualifications? Another possible interesting group of persons are the employees of the company. How well are the employees satisfied with their work at the company? Do they, for example, complain about working in poor working conditions? A less obvious group of persons that can be searched for to reveal possible interesting information about important persons working at the company are the family members and friends of these persons. Family and friends might give an impression about someone's behavior and therefore also might be useful.

3. Sector/Environment. The location of the company contributes to its success or failure. For example, a company is more likely to succeed whenever it is located in an area with many potential customers and few competitors. Therefore, information about the environment of the company might be useful. Furthermore, the performance of the industry sector the company belongs to might also be useful for estimating a company's creditworthiness.

5.5 Information sources

An information source is considered to be a website on which information about the company or persons related to the company can be found. This information can be either data that is explicitly present on these websites, or data that is implicitly present and can be derived from this explicit data. The latter is called metadata. An example

of metadata is the frequency of and the time between posts or updates. The information sources discussed in this section follow from the results of the literature study and the survey.

5.5.1 Social Media

Social media are platforms on which people can publish and access information, collaborate on a common effort, or build relationships [54]. Some social media websites (e.g., LinkedIn and Facebook) make a distinction between user profile pages and company profile pages in which the latter allows the company to fill in specific company related fields. The information on these platforms can be publicly shared or restricted to connections of the account's owner. Often, the public account only shows a limited amount of information that the account's owner has published. Some social media allow account owners to grant access to their private profile to non-connected users. Although this results in more information about the person, asking for explicit access from user to their private account is not considered in this research because of issues discussed in Section 1.2 and the loss of (reference) uncertainty. Some indicators that could be extracted from social media were already discussed in Section 5.2.3.

5.5.2 Company Website

From the results of the survey in Appendix E it became clear that some persons are interested in a company website. However, it did not become clear in which specific information on these websites they are interested. The information found on company websites can differ largely. The website of Shell[9] was visited to get an idea of what information can be found on a good corporate website. The website of this company was chosen since it was rated best on corporate online effectiveness by Bowen Craggs & Co[10]. Several sections that often occur on corporate websites can be recognized on this website. Depending on the type of company (Business-2-Business vs. Business-2-Consumer), the industry it is active in and the size of the company the corporate website can have one or several of the following sections:

- Home page. The home page is the main page of the website and serves to show the visitor the most important information available on the website and allow them to navigate to other pages.

[9] http://www.shell.com
[10] http://www.bowencraggs.com/FT-Bowen-Craggs-Index/Interactive-results-table

- About us. The "about us" section of the website gives and informative description of the company. Information that often is discussed within this section is:
  - History
  - Mission
  - Operations
  - Biographical information on the founders, board members and sometimes also information on all the other employees.
  - News / Blog
  - Clients, suppliers, achievements and project partners
  - Jobs and Careers
- Products and services. The products and services section of the website describes the (major) products and services offered by the company. This can consist both of an informal sales description or a more formal specification.
- Events / Calendar. This section shows upcoming and/or past events which are organized or sponsored by the company; or in which the company otherwise participates.
- Contact information. The contact section lists addresses, phone numbers and or e-mail addresses.
- Store locator. The store locator can be used to find nearby locations of the company or of companies where their products or services can be found.
- Frequently Asked Questions (FAQ) / Help. The help section gives answers to the most common question related to the company or its website.
- Terms and conditions. The terms and conditions describe the website's content and how visitors are allowed or not to use it.
- Privacy policy. The privacy policy describes which personal information is collected from visitors to the website.

Several possible creditworthiness indicators that could be extracted from a company website were identified:

- Although it might not always be trivial to detect, the presence of these elements on a corporate website could perhaps be used as an indicator. In addition, the amount of words within these sections could be counted and perhaps be used as an indicator.

- Another indicator that can possibly be extracted from a corporate website is the level of mobile friendliness. Google offers a tool that can check whether a website is mobile friendly[11]. The idea behind this is that a website that is accessible well on various devices has a larger potential audience. It shows that more attention is paid to the communication with potential customers. However, care must be taken when using this as an indicator, because it some types of companies (e.g., IT companies) might be more likely to have a mobile device optimized website than others.

- From the survey it became clear that the update interval could perhaps also be interesting as creditworthiness indicator. A higher update interval could, for example, reflect higher business activity. A method to determine the update interval of a website could involve The Wayback Machine of The Internet Archive[12]. Limitations of this method are that the interval is only an approximation, because websites are not crawled daily. In addition, often only the home page is crawled making it only possible to detect changes there.

5.5.3 Statistics Bureau

Many countries have a statistics bureau, which is responsible for collecting, processing and publishing statistics on behalf of the government and businesses. The themes for which figures are published often include: general and regional statistics; economy and finance; population and social conditions; industry; trade and services; agriculture; forestry and fisheries; international trade; transport; environment and energy; and science and technology. An example of such statistics bureau is the Central Bureau of Statistics (CBS) in the Netherlands, which makes their figures accessible via the electronic databank StatLine [13].

The added value of this online data can be put in doubt since ratings often already include some of these figures into their calculation (See section 2.1.3). We will not focus on this information source in this research, because the data provided by the statistics bureau does not have data uncertainty and therefore is less relevant for this research.

5.5.4 News Items

News items published on news websites could perhaps reveal information on how creditworthy a company is. The amount of news items that can be found and the subjects

[11] https://www.google.com/webmasters/tools/mobile-friendly
[12] http://archive.org/web/
[13] http://statline.cbs.nl

of these items largely depends on the size of the company. Several news item subjects related to companies can be identified:

- Company performance, such as turnover figures; profit and loss figures; number of layoffs and hires; expansion or downsizing details; and bankruptcy rumors or details.
- (New) Products and/or services.
- Working conditions and strikes.
- Customer satisfaction and prizes awarded to the company or its key persons.
- Actions taken or (rumored) to be taken by the competition that might impact the company's performance.
- Events organized or sponsored by the company or in which it otherwise participates.

However, text analysis is a study on its own and is out of scope of this research. No existing off-the-shelf solutions text analysis tools were found. For example, sentiment analyzers are not very useful for news articles because these tend to be objective. Therefore, news items are left out of scope of this research.

5.5.5 Google Reviews

One survey participant suggested using Google reviews as a creditworthiness indicator. Google allows people to review companies via Google+ by giving a star rating and a description of their experience with the company. The star rating is an indicator that could relatively easily be extracted. Google sometimes incorporates the ratings of a company from another website into its search result. These could possibly also be used.

5.5.6 Google Trends

Google Trends was also suggested in the survey as possible creditworthiness indicator. Google Trends is a tool that analyses part of the Google search queries to determine how many searches are done on those terms compared to the total amount of search queries in that period[14]. Users can compare these trends per region or per period. If too few people search for a certain term, this term is excluded from Google Trends.

[14]https://support.google.com/trends/answer/4355213?ref_topic=4365599

5.5.7 Job Vacancy Website

The amount of job vacancies might indicate how well business is going for a company. Several job vacancy websites exist in the Netherlands. We only considered three well-known websites: monsterboard.nl; nationalevacaturebank.nl and indeed.nl. Besides the kind of job, these websites also allow to search by employer name which allows us to determine how many job vacancies a company has outstanding.

5.6 Information availability and usability

Although credit lenders are interested in certain online information, it might not be publicly available and therefore not eligible. Availability of information is important, because the data mining tool needs sufficient samples for training. Therefore, the availability of information was determined for companies for which a rating was collected in Chapter 4 using a sample of 50 subjects. The results of these availability tests are used to determine which information and which information sources can be used for testing the added value of accounting for uncertainty. Table 5.1 summarizes the results of the availability tests.

TABLE 5.1: Availability of online information.

Source	**Subject**	**Percentage found**
LinkedIn	Director / manager / contact	84%
	Company	47%
Twitter	Director / manager / contact	12%
	Company	44%
Facebook	Director / manager / contact	14%
	Company	45%
Google Reviews	Company	6%
Google Trends	Company	25%
Job Vacancy Websites	Company	24%

5.6.1 Social Media

This subsection discusses the results of a test that was performed to determine the presence on social media of companies from our data set and persons related to these companies. In addition, to determine which specific information is published and publicly available, a more specific test is performed for the social media websites on which the companies / persons are considerably present.

Presence of persons / companies on social media

In total, 50 directors / managers / contacts of several random companies from our data set were searched on LinkedIn, Twitter and Facebook. The search was done while being logged in on these websites. A person was marked found if there was sufficient evidence that a found account actually belongs to the person that was searched for. As can be seen in Table 5.1, a large amount of company directors / managers / contacts could be found on LinkedIn. However, the percentage of persons found on Twitter and Facebook is quite low. The explanation for this is twofold:

- Often, users protect their information from non-connections. Furthermore, some users that do have an account did not fill in all information on their account. This makes it sometimes impossible to determine if the account actually belongs to the person that was searched for. Because the amount of information that is available in these cases is very limited, the profile often also is not very useful as information source. For example, some Twitter accounts that were found only contained one or a few Tweets which were posted around the time the account was created. After that, the person did not use the account anymore.
- The names of the directors / managers / contacts in the data set are legal names. The legal first name often deviates from a person's nickname. For example, the nickname of a certain person with the legal first name 'Gerardus' was 'Gertjan'. In the manual search, mainly the last name in combination with the company name was used as a search criterion, because these nicknames are not always obvious or the same for a certain legal name. This strategy worked well for LinkedIn, because most LinkedIn accounts contain information on the company the user currently works. However, this information often is not publicly available on Twitter or Facebook.

Based on these results, LinkedIn seems most promising to use as a social media source when searching for persons related to a company. However, note that the low percentage found on Twitter and Facebook does not necessarily mean that these persons actually do not have an account. With the limited information available, both on the profile and in the data set, we were not able to find the user. In some cases, the legal company name also substantially differs from the commonly used company name, making the search for information about it more difficult.

In addition to this, a study was done on the amount of companies with a profile on LinkedIn, Twitter and Facebook. Again, around 50 random companies were selected from the data set and were searched on these social media websites. The results of

this test can also be found in Table 5.1. The results suggest that it is worth using the company pages on these websites, because a considerable amount of around 45% of the companies in our data set have a profile on LinkedIn, Twitter and Facebook. For some companies, LinkedIn and Facebook have generated a page automatically. Although from a search perspective these pages might be correct (the page refers to the company that was searched for), these pages were nevertheless classified as invalid. This was done because these pages contain only information that was already known from the data set, such as website URL and company address.

Information available on profile

For personal LinkedIn accounts, and for corporate LinkedIn, Twitter and Facebook accounts, a small sample of 50 companies from our data set was taken and the specific information available on these profiles was determined. Table 5.2a and 5.2b show the information availability results of personal and corporate social media profiles respectively. We will now shortly summarize which information seems, based on this availability, promising to use as an indicator.

For personal LinkedIn accounts, based on the availability it seems promising to use as an indicator: the number of connections, the number of groups the person is registered to, the number of members the person is following, the presence of a profile photo and the number of skills. More specifically, for skills, the number of skill endorsements could perhaps also be used as an indicator. Also the specific type of skills might be an interesting indicator. However, it is more difficult to train a classifier on all possible skills, given the large variety of skills that a person can place on his or her profile. The listed experiences can be used, for example, to determine the number of job switches. The same could be done for listed educations. In addition, education could be used to extract the level of education (e.g., Middelbaar beroepsonderwijs, Hoger beroepsonderwijs, Universiteit) and use this as an indicator.

For corporate LinkedIn pages, the number of followers the profile has, the number of employees and the amount of recent updates posted might be eligible as a creditworthiness indicator. For corporate Twitter profiles, the number of followers, the number of tweets posted by the company, the number of accounts the company follows, whether the account contains photos / videos and the number of tweets that have been 'favorited' seem to be a promising starting point to use as an indicator. In addition, based on the literature discussed in Section 5.2.3, also the level of engagement of the company and the sentiment of the tweets received could be used. For Facebook, the number of likes received by the company, the number of posts on the profile, whether the profile contains

TABLE 5.2: Information available on personal/corporate social media profile.

(A) Personal user profile

Source	Attribute	% Found
LinkedIn	Full name	100%
	# Connections	100%
	Title	90%
	Experience	90%
	Location	88%
	Current position	88%
	Industry	86%
	Following	79%
	Groups	60%
	Past position	57%
	Profile photo	55%
	Skills	55%
	Education	55%
	Languages	26%
	Summary	21%
	Recommendations	21%
	Interests	5%
	Organizations	5%
	Contributions	2%
	Projects	2%
	Diploma / certifications	2%
	Voluntary work	0%
	Awards	0%

(B) Corporate user profile

Source	Attribute	% Found
LinkedIn	Company name	100%
	#Followers	100%
	Website URL	100%
	Industry classification	100%
	Company logo	91%
	About us	91%
	Business type	91%
	Size	91%
	List of employees	87%
	Headquarter location	83%
	Specialism	78%
	Founding date	70%
	Recent updates	57%
	Interesting updates	48%
	Commented updates	35%
	Careers	17%

Source	Attribute	% Found
Twitter	Company name	100%
	Account name	100%
	# Followers	100%
	Profile image	96%
	Tweets	96%
	Replies	96%
	# Follows	91%
	Website URL	91%
	Business location	87%
	Registered on	87%
	About company	83%
	Photos & videos	78%
	Favorites	52%
	Lists	22%

Source	Attribute	% Found
Facebook	Company name	100%
	# Likes	96%
	Profile image	92%
	Posts	92%
	Website URL	88%
	Photos	88%
	About us	83%
	E-mail	83%
	Address	71%
	Phone number	67%
	# Visits	58%
	Founding date	50%
	Reviews	46%
	Events	42%
	Products / services	38%
	Videos	29%
	Awards	21%
	Mission	21%
	Milestones	17%

photos / videos and the number of visits to the company according to the profile, seem to be a good starting point to use as an indicator for corporate Facebook pages.

5.6.2 Google Reviews

Table 5.1 shows that of 50 randomly chosen companies from the data set, only 6% were found to have a Google review. This does not seem to be an information source that can be used for many companies. When searching for reviews, we noted that more company reviews were available within the Google search results. Most of these reviews originated from the telephone directory website 'telefoonboek.nl'. A review was incorporated within the Google search results for 20% of the companies. Although this is higher, it is still quite low.

5.6.3 Google Trends

A test was performed to determine how many companies could be found on Google Trends. Search queries only show up in Google Trends if they have enough search volume. No information was found on what this minimum amount of volume is. In total, 50 random companies from the data set where searched on Google Trends. At first, the entire legal company excluding the business type (e.g., B.V.) was used as a query. In this case, as can be seen in Table 5.1, 25% of queries returned a result. Although using only subsets of the company name increased this percentage to 67%, it includes many non-related search queries, e.g., due to the query consisting of some general terms or a common last name.

5.6.4 Job Vacancy Website

Table 5.1 shows that a job vacancy was found on at least one of the three websites for 24% of 50 randomly chosen companies from the data set.

5.6.5 Others

- **Entrepreneur behavior**. The existing solution to extract the behavior of an entrepreneur from social media that was discussed in Section 5.2 uses Twitter. The availability study suggests that directors / managers / contacts of the company are difficult to find on Twitter. Hence, this does not seem to be a promising indicator.

- **Regulatory burden**. Although the regulatory burden could perhaps be extracted, it cannot be considered as uncertain data and therefore is of less importance for this research.

- **Living area / Business location**. Social media could be used to find the living area of someone. Only a reasonable amount of directors / managers / contacts of a company could be found on LinkedIn, which only shows a rough location (e.g., Amsterdam area). This makes it difficult to classify the person living in a good or underprivileged neighborhood, because this differs within a city and certainly within a larger region. For companies, it is possible to classify the neighborhood because their address is known from the data set, however, we also lose uncertainty.

- **Other information subjects**. With the data available in the data set it is hard (or impossible) to determine who the other financiers and past owners are. Suppliers and customers could perhaps be derived from the connections of a company on social media or from the people / companies who posted a message on the company's profile. Employees of a company could be derived from LinkedIn, however, this requires a paid premium account. Before we can use these social media accounts of employees as an indicator, we first need to determine if an individual social media account can be used as an indicator. The same holds for friends and family of the company owner. Other companies within the same sector might be determined using the NACE industry classification, however, these classifications are often quite broad. Furthermore, before a corporate social media account is eligible for this research, it requires testing whether a single corporate social media account can be used as an indicator.

5.7 Collecting data

Based on the results of the survey, literature and the results of the availability tests in the previous section, we decided to use personal LinkedIn accounts and corporate Twitter and Facebook accounts as information sources. Note that, although the availability of corporate LinkedIn accounts was roughly the same as corporate Twitter and Facebook accounts, it is not used in this research because of time-constraints. Corporate LinkedIn accounts were chosen to be left out, because in that case we still have information from all three social media platforms and both information of persons and companies. Before the online information of these sources can be used, we need to harvest the data. For determining the performance of online information as predictor of a company's creditworthiness, the model is trained and tested on data for which uncertainty is not modeled. A manual search for the URLs of LinkedIn owner accounts, Twitter corporate

accounts and Facebook corporate accounts was done for a sample of companies/persons from our financial data set to obtain a set of online information without (reference) uncertainty. The results of this manual search are discussed in Section 5.8. In this section, we explain how the information was harvested from the selected information sources.

5.7.1 LinkedIn user profile

Although LinkedIn has an API for accessing full profile information of LinkedIn members, it could not be used for this research. Access to the API needs to be requested and is only granted by LinkedIn to companies that use it for their career site[15]. As an alternative, the web harvester tool Import.io[16] was used to extract information from LinkedIn pages. This tool allows developers to create an API that harvests data of a website when it is called. Training of the web harvester is done through a simple point and click interface.

The information visible on public LinkedIn profiles (i.e., profiles that are visible for anyone who searches the person on Google, Yahoo!, Bing etc.[17]) is often limited. However, generally, a LinkedIn profile is fully visible to LinkedIn members who have signed in to LinkedIn. Although Import.io has some functionality to train the web harvester to first log in onto a website before harvesting the data, it was not compatible with LinkedIn. Therefore, a proxy has been developed that retrieves requested LinkedIn profiles while being logged in as a LinkedIn member. Import.io requests LinkedIn profiles through this proxy. This is illustrated in Figure 5.4.

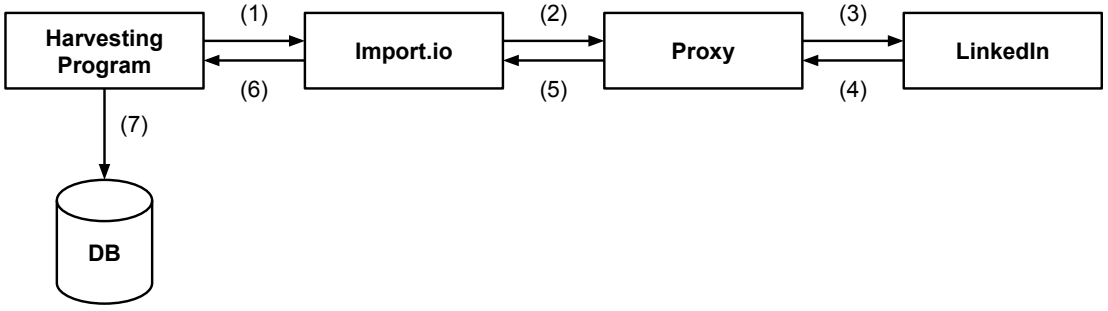

FIGURE 5.4: Flow of harvesting LinkedIn profile data.

Table 5.3 shows the data that was harvested from LinkedIn together with the indicators that have been derived from this data.

[15] https://developer.linkedin.com/docs/apply-with-linkedin
[16] http://import.io
[17] http://help.linkedin.com/app/answers/detail/a_id/77

TABLE 5.3: Data harvested from LinkedIn together with indicators derived from this data.

Profile Data	**Indicator**	**Data Type**
#Member connections	#Member connections	Numeric
Skills	#Skills	Numeric
	#Edorsements	Numeric
Experiences	#Months experience	Numeric
	#Job switches	Numeric
Educations	#Education switches	Numeric
	Education level	Numeric: 0 (none), 1 (basis), 2 (voortgezet), 3 (MBO), 4 (HBO), 5 (universitair)
Summary	Presence of summary	Nominal: true, false

5.7.2 Twitter corporate profile

Twitter has a REST API[18] that can be used to harvest data from Twitter profiles. The Twitter4J Java library[19] was used to communicate with this API. Table 5.4 shows the indicators that are derived from the harvested Twitter data.

TABLE 5.4: Data harvested from Twitter together with indicators derived from this data.

Profile Data	**Indicator**	**Data Type**
#Tweets	#Tweets	Numeric
#Followers	#Followers	Numeric
#Favorites	#Favorites	Numeric
#Friends	#Friends	Numeric
Photos & videos	#Media	Numeric
Verified	Verified	Numeric

5.7.3 Facebook corporate profile

Facebook has a Graph API[20] that can be used to harvest data from Facebook profiles. The RestFB Java library[21] was used to communicate with this API. The Facebook API does not allow retrieving Facebook reviews when no permission is granted by the owner of the profile. Hence, it is not included as indicator. Table 5.5 shows the indicators that are derived from the Facebook data that could be harvested.

[18] https://dev.twitter.com/rest/public
[19] http://twitter4j.org
[20] https://developers.facebook.com/docs/graph-api
[21] http://restfb.com

TABLE 5.5: Data harvested from Facebook together with indicators derived from this data.

Profile Data	**Indicator**	**Data Type**
#Talking about	#Talking about	Numeric
#Were here	#Were here	Numeric
#Likes	#Likes	Numeric
Events	#Events	Numeric
Milestones	#Milestones	Numeric
Videos & photos	#Media	Numeric

5.8 Result

Table 5.6 shows the amount of manually searched personal LinkedIn accounts and corporate Twitter and Facebook accounts. The persons / companies of this data set were randomly picked from the 3579 companies of the in Chapter 4 explained data set. The approximate 65% of company owner LinkedIn accounts found in this sample is less compared to the 84% found in Chapter 5 (Table 5.1 in Section 5.6.1). However, the approximate 52% of corporate accounts found on both Twitter and Facebook in this sample is slightly higher than the 44% and 45% respectively. Furthermore, note that there are more LinkedIn entries than Twitter and Facebook entries, because several companies have multiple owners.

TABLE 5.6: Manually constructed data set of social media accounts.

Source	**#Found**	**#Not Found**	**#Total**
LinkedIn	284	152	436
Twitter	202	185	387
Facebook	203	184	387

If the PD of companies in our data set would deviate much per industry sector, it might be difficult to make a generalized model that can predict the creditworthiness of companies in all sectors. Therefore, we checked how much the average PD of companies deviate per industry sector. Table 5.7 shows the average, minimum and maximum PDs per sector for companies in our data set. The average PD for most sectors is around the 4%, except for the sectors 'Textiles, wearing apparel, leather', 'Publishing, printing' and 'Wood, cork, paper' of which only a few companies are within our data set. It therefore does not seem likely that a difference in average PD per industry sector will form a problem in constructing a model.

Table 5.7: PD of companies in our data set per industry sector.

Sector	**Amount**	**Avg PD**
Chemicals, rubber, plastics, non-metallic products	16	1.73021%
Construction	27	3.67011%
Education, Health	16	4.83131%
Food, beverages, tobacco	15	3.20112%
Gas, Water, Electricity	3	2.61383%
Hotels & restaurants	6	0.01576%
Machinery, equipment, furniture, recycling	25	5.93590%
Metals & metal products	7	3.61502%
Other services	103	3.55360%
Post & telecommunications	6	4.60545%
Primary sector	16	2.49327%
Publishing, printing	5	9.07296%
Textiles, wearing apparel, leather	2	12.02384%
Transport	27	5.35877%
Wholesale & retail trade	112	4.24710%
Wood, cork, paper	1	0.23114%

5.9 Conclusion

In this chapter, we first studied which information is possibly interesting as a creditworthiness predictor using a survey and a literature study. The results of the survey show that, according to the participants, the LinkedIn, Twitter and Facebook social media websites can be useful for making creditworthiness predictions and that information about the company in question and its owner(s) is most relevant. Literature on related existing solutions, on social media as business intelligence and on language analysis in tweets confirm the potential of social media websites as an information source for creditworthiness predictions. Tests were performed to determine the fraction of companies / owners that can be found in an information source and which particular information is available within these sources. In total, for 84% of the company owners in our sample a personal LinkedIn profile was found. For 44% and 45% of the companies in our sample a corporate Twitter and Facebook account was found respectively. From the results of these tests, we concluded that these sources can be used for constructing a prediction model. Based on the information availability within personal LinkedIn accounts, we decided to extract as an indicator: the number of connections, skills, endorsements, months experience, job switches, education switches; and the education level and the presence of a summary. In particular numerical indicators expressing the volume of posts, influence and engagement were extracted for corporate Twitter and Facebook social media accounts. The setup used to harvest information from these sources was explained. We explained that, because model construction is done on data for which

uncertainty is not modeled, we performed a manual search for a sample of companies / persons of our data set from Chapter 4. The result was a data set consisting of 387 companies and 436 persons related to these companies for which the online information was harvested and the indicators were extracted.

Chapter 6

Indicator performance as creditworthiness predictor

In this chapter, we explain how creditworthiness prediction models were created using the online data of Chapter 5 as indicators and the financial data of Chapter 4 as target variable of the model. Two experiments are performed. In the first experiment, we test the prediction performance of online data using only the indicators extracted from this online data. We explain how the data was preprocessed and discuss the results. In a second experiment, we test whether the online information can be used for making predictions of a company's creditworthiness when it is used in combination with financial data. The results of these experiments, allows us to give answer to the third subquestion on the performance of indicators extracted from online information as predictor for creditworthiness.

6.1 Experiment 1: Accuracy of a model using Internet features only.

In this experiment, the performance of the selected Internet features as a predictor of a company's creditworthiness is tested using data mined rating models. To test the performance of online information as creditworthiness indicator, a model needed to be constructed. The Waikato Environment for Knowledge Analysis[1] (WEKA) data mining tool is used to construct this model. The rating models are constructed using the classifiers: Naive Bayes, Support Vector Machine (SVM), J48 (C4.5) and Random Forest. These schemes are run 10 times with 10-fold cross validation. Figure 6.1 illustrates the

[1] http://www.cs.waikato.ac.nz/ml

idea behind this experiment. The accuracy of the models is determined by comparing the rating predicted by these models to the URA rating, which have been computed based on financial data only. The performance of the classifiers is compared to the performance of the model constructed using the ZeroR classifier. This classifier is often used as a baseline and ignores all input attributes and simply predicts the most frequent output class. Comparison to this baseline is done using a two-tailed paired t-test with a confidence of 5%. Refer to Appendix C for more background information about the classifiers used within this experiment.

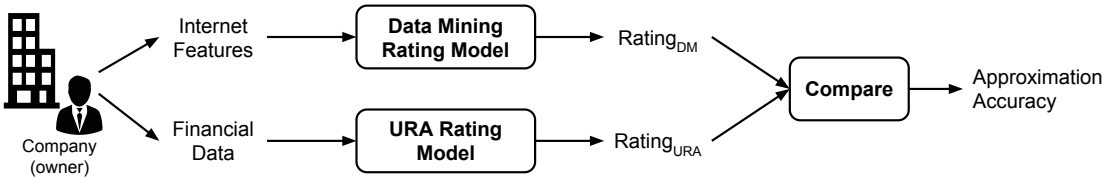

FIGURE 6.1: Approach to determine the performance of Internet features as creditworthiness predictor.

6.1.1 Preprocessing data

The data collected was preprocessed before training and testing the model. This is discussed in the following subsections.

Missing and multiple indicator values

Indicators were assigned zero values in case of the absence of those indicators on a social media profile. In case no social media account was found at all, all the corresponding indicators were assigned zero values. In case a company has multiple owners, the numerical indicators were averaged and Boolean values were ORed.

Rating classes

While performing some initial tests, it was found that, given the limited amount of Internet features, the task of predicting all 19 possible rating classes is too comprehensive for the classifier. Furthermore, as can be seen in Figure 6.2, the amount of instances for some rating classes is quite low. Because the test will run in cross-validation mode, even fewer instances of these classes will be present in each run of this cross-validation, making it hard for the classifier to train on these classes. Therefore, it was decided to reduce the number of possible output classes to 6: strong, satisfactory, adequate, uncertain, risky and inadequate. These classes are based on the description of the URA

ratings in Table 2.1 in Chapter 2. Figure 6.3 shows the distribution of the companies in the data set over these 6 ratings. Compared to Figure 6.2, more ratings per rating class are available.

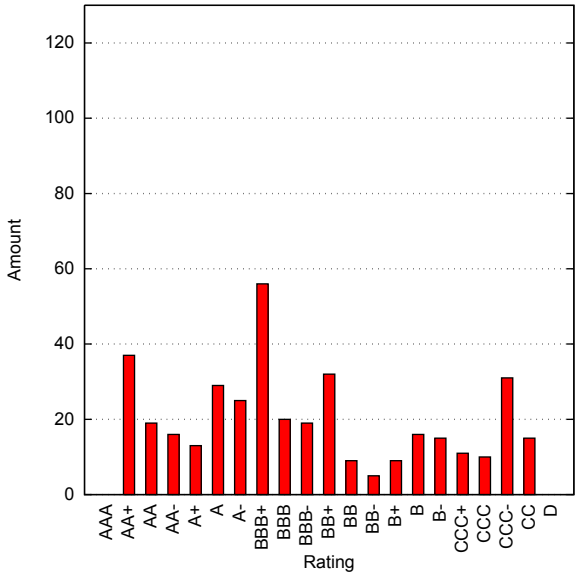

FIGURE 6.2: Original rating distribution of the data set.

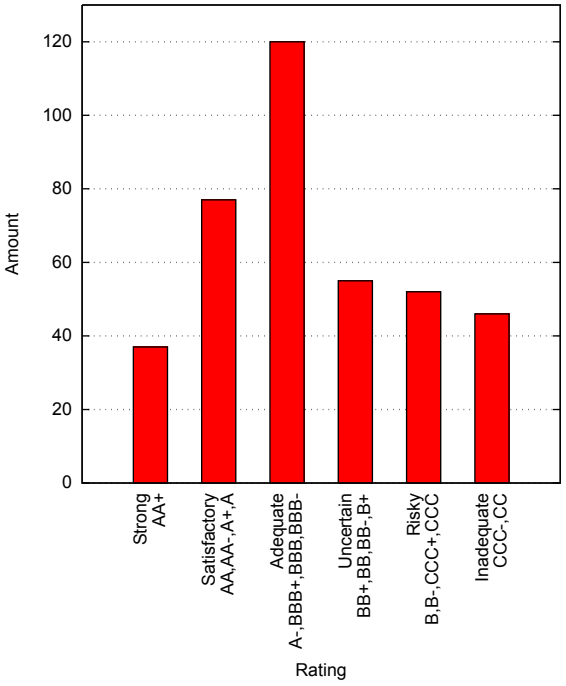

FIGURE 6.3: Rating distribution of the data set with a reduced number of rating classes.

Most Internet features that are used within this research are quite static over time. However, ratings differ from year to year. To rule out that this change in rating is a major issue for our sample, a check was done on the fluctuation in rating. Figure 6.4 illustrates this fluctuation by showing the percentage of ratings that has increased or decreased a certain amount of steps in one and five years respectively. As can be seen, most ratings have only changed slightly between 0 and 2 steps after both 1 and 5 years. Ratings on the reduced rating scale are an aggregate of at most 4 original rating classes. Therefore, if the rating has changed, in most cases it will be changed at most a single step on the reduced rating scale.

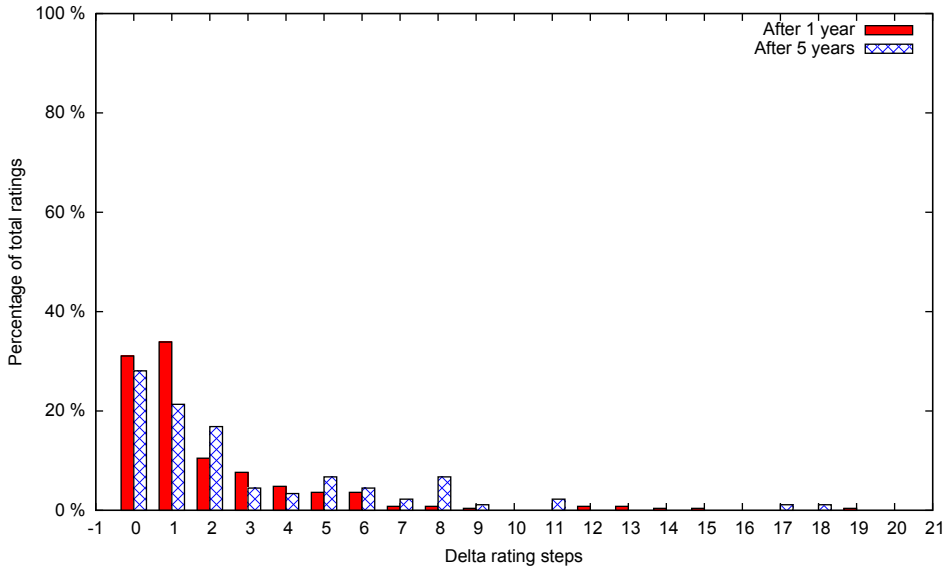

FIGURE 6.4: Fluctuation in ratings after 1 and 5 years.

Discretization

According to [55], discretization of numeric attributes is not only essential for learning schemes that can only handle categorical data. Despite some loss of information when discretizing attributes, often even schemes that can handle numeric attributes produce better results when the attributes are discretized. Therefore, an unsupervised discretize filter was used to discretize the numeric values of each Internet feature into 10 equal-frequency bins.

Attribute Selection

Often, it is the case that some included attributes are redundant. Although some learning schemes try to only select relevant attributes (e.g., decision trees), often there are still too many attributes that a classifier can handle [55]. In practice, the performance can

often be improved by only selecting most relevant attributes. This process of automatically searching for an optimal subset of attributes is called attribute/feature selection. The CfsSubset attribute evaluator which "evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them."[2] was used to perform attribute selection. The Internet features that were selected by this evaluator are shown in Table 6.1.

TABLE 6.1: Internet features selected by the CfsSubset attribute evaluator.

Source	**Feature**
LinkedIn	#Endorsements
	#Months experience
	#Job switches
	#Education switches
Twitter	#Followers
	#Favorites
	#Media
Facebook	#Were here
	#Events
	#Milestones
	#Media

6.1.2 Results

The results of experiment 1 can be found in Table 6.2. The table shows the result of the paired t-test in which models 2-5 are paired to a ZeroR baseline model (1).

TABLE 6.2: Paired T-Test comparing prediction accuracy of Internet feature model

Data set	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Internet features (selected)	31.01	21.02 •	27.34 •	26.15 •	31.01

◦, • statistically significant improvement or degradation

(1) ZeroR
(2) Naive Bayes
(3) J48
(4) Random Forest
(5) SVM

For each model constructed using the selected Internet features and tested classifiers, the percentage of correctly classified instances is shown (i.e., how often does the rating predicted based the selected Internet features matches the rating calculated over the

[2] http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html

financial data using the URA rating model). As can be seen in the table, all models have an accuracy of approximately 30%. The model constructed using the Naive Bayes, J48 and Random Forest classifiers perform statistically worse than the model constructed using the ZeroR classifier. The model constructed using the SVM classifier yields the same accuracy as the ZeroR baseline model. Studying the SVM model showed that it, just like the ZeroR model, would always predict the adequate rating class. These results suggest that the selected attributes that were extracted from an online source cannot solely be used to make a prediction of a company's creditworthiness.

6.2 Experiment 2: Accuracy of a model which combines Internet features with financial data.

Because none of the models in experiment 1 performed better as the ZeroR baseline model, a second experiment was performed in which the effect of extending a data mined financial model with Internet features is tested. The approach of this experiment is illustrated in Figure 6.5. The idea behind this experiment is that the Internet features by itself might not be good predictors, but perhaps are in combination with financial indicators. In this experiment, the rating that was computed over the financial data using the URA rating model is again used for validation. First, this URA rating model is approximated by creating a data mining model using the same financial data used by the URA rating. The accuracy of this approximation is used as a baseline. Then, this financial model is extended with Internet features. The accuracy of this extended model as approximation of the URA rating model is determined. Both approximation accuracies will be compared to determine whether these Internet features have added value as creditworthiness predictors. Comparison again is done using a two-tailed paired t-test with a confidence of 5%.

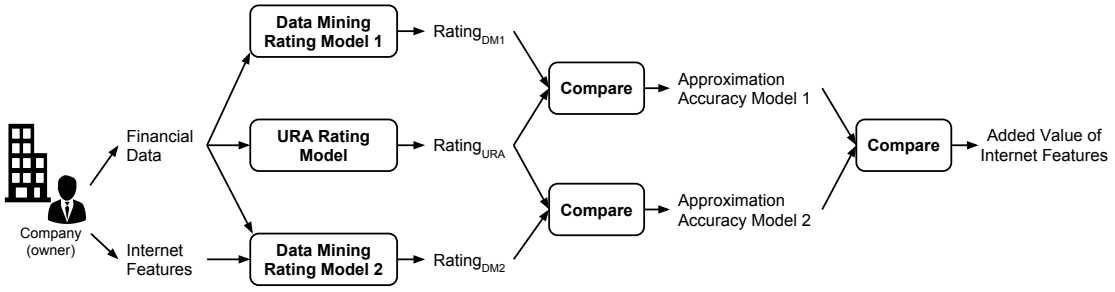

FIGURE 6.5: Approach to determine the performance of Internet features when combining them with a financial model.

6.2.1 Preprocessing data

The financial data necessary to perform this experiment was extracted from the XML documents described in Section 4.3.1 of Chapter 4. The same set of companies as in experiment 1 is used. This experiment was again performed with the reduced number of 6 rating classes. The Internet features were discretized into 10 equal-frequency bins. The financial data was not discretized, because it was not found to make a significant difference while running some initial tests and, as far as we know, the original URA rating model also does not discretize its data. Attribute selection was again performed as a preprocessing step. The financial features that were selected by the CfsSubset attribute evaluator are:

1. Profit loss before tax
2. Borrowed funds earnings before interest, taxes, depreciation and amortization
3. Financial leverage
4. Earnings before interest, taxes, depreciation and amortization margin
5. Profit before taxes / total assets
6. Return on assets
7. Trade and other payables current

When adding also the Internet features to this financial data, the attribute evaluator in addition selects the #Month experience (LinkedIn) as an attribute.

6.2.2 Results

Table 6.3 shows the result of a paired t-test in which models 2-5 are paired to a ZeroR baseline model (1). For each model constructed using the selected financial data and tested classifiers, the accuracy (i.e., the percentage of correctly classified instances) of the financial data mining model as predictor of the original URA rating is shown. As can be seen, Naive Bayes and SVM do not perform statistically better or worse than the 31% accuracy of the ZeroR classifier. However, the J48 and Random Forest classifiers do perform statistically better than the ZeroR classifier with an accuracy of 63% and 68% respectively.

The models including Internet features were constructed using the two best performing classifiers in our financial model (J48 and Random Forest) to test whether these features have added value as creditworthiness predictors in combination with the financial data. The results of a paired t-test for both classifiers in which the models constructed using the data sets including Internet features are paired to the data set consisting of only financial

TABLE 6.3: Paired T-Test comparing prediction accuracy of the financial model

Data set	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Financial data (selected)	31.01	33.43	62.77 ∘	67.99 ∘	31.01

∘, • statistically significant improvement or degradation

(1) ZeroR
(2) Naive Bayes
(3) J48
(4) Random Forest
(5) SVM

data is shown in Table 6.4. The table again shows the percentage of correctly classified instances. As can be seen in the table, both models have an accuracy of approximately 63% after adding the selected Internet feature (#Months experience). The accuracy of the Random Forest model decreases significantly when adding this feature, while the accuracy of the J48 model increases slightly when adding the feature. Although the attribute evaluator selected only one Internet feature, in general adding only a single extra feature is not likely to improve accuracy much. Therefore, we also studied the effect on accuracy when adding all collected Internet features to the financial model. As can be seen in the table, for both classifiers the accuracy decreases when additionally adding the other Internet features as well. These results suggest that these Internet features also have no added value when combined with financial data.

TABLE 6.4: Paired T-Test comparing prediction accuracy of the model with and without Internet features

Data set	**(1)**	**(2)**
Financial data (selected)	68.17	63.10
Financial data (selected) + Internet features (selected)	62.39 •	63.93
Financial data (selected) + Internet features (all)	61.46 •	59.47

∘, • statistically significant improvement or degradation

(1) Random Forest
(2) J48

6.3 Conclusion

In a first experiment, the performance of Internet features as a predictor of a company's creditworthiness was tested using solely these Internet features. The Internet features that were selected while preprocessing the data are: #Endorsements, #Months experience, #Job switches, #Education switches, #Followers, #Favorites, #Twitter media,

#Were here, #Events, #Milestones and #Facebook media. The accuracy of models constructed using the Naive Bayes, SVM, J48 and Random Forest classifiers were compared to a ZeroR baseline model which always outputs the rating class it found to be most occurring. None of these classifiers performed significantly better as the 31% accuracy of this baseline. These results show that in this setup the selected Internet features cannot be used solely to make a prediction of a company's creditworthiness.

Because Internet features might have predictive value when used in combination with financial data, a second experiment was performed. In this experiment, models were created using solely financial data and using both financial data in combination with selected Internet features. The performance of both models was compared by determining their approximation accuracy to the URA rating. The two best performing financial models had an approximation accuracy of 68% and 63%. Adding Internet features to these models gave mixed results with a significant decrease and an insignificant increase in approximation accuracy. From this can be concluded that the selected Internet features do not have an added value for predicting a company's creditworthiness in this setup. Several research directions are proposed in Section 9.2 of Chapter 9 which might result in a more accurate model for predicting a company's creditworthiness using online data.

Chapter 7

Including uncertainty of online data

In the previous chapter, a prediction model was trained and tested using a manually created data set for which it was certain that the online data of the persons and companies in that data set was related to the company of which the rating was predicted. However, in practice, automatically searching for information often is accompanied with uncertainty. In [15], Van Keulen et al. claim that dealing with alternatives (i.e. possible worlds) along with their confidences might yield better results. This chapter discusses an experimental setup for testing the effect on accuracy of applying the possible worlds concept to the creditworthiness prediction model that uses online data. A search system is described to find candidate results for an entity and to determine the level of (un)certainty that a result belongs to the searched entity. Lastly, we explain two approaches for dealing with uncertainty in data mining.

7.1 Setup

This section describes a setup for testing how incorporating uncertainty affects the accuracy of our creditworthiness prediction models. In this research, we focus on reference uncertainty, because the online data identified in Chapter 5 has no data uncertainty. The setup is illustrated in Figure 7.1. The idea is to test the accuracy of the data mining rating model using two different sets of data. In one case, only Internet features of the best matching result are used for making predictions. In the other case, the Internet features of multiple alternatives are used together with the corresponding level of (un)certainty (i.e., probabilities expressing how well the alternative results match the details known about the subject for which the rating is predicted) for making prediction. In both

cases, the predicted ratings by the data mining model can be compared to the URA rating to determine the prediction accuracy. In turn, these prediction accuracies can be compared against each other to determine whether the prediction accuracy improves when uncertainty of the online data is accounted for. The dotted lines indicate that the experiment can be done both with the model using solely online data (as in Section 6.1 of Chapter 6), or the model using both online data and financial data (as in Section 6.2 of Chapter 6).

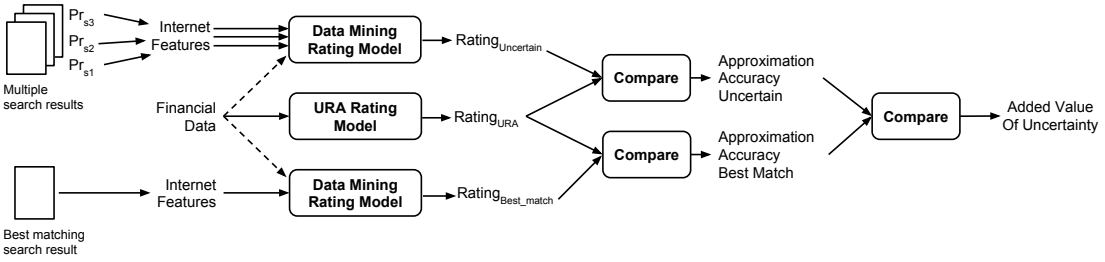

FIGURE 7.1: Testing the effect of uncertainty on prediction accuracy.

The described setup requires a system for searching candidate results that in addition can determine the level of reference uncertainty of these results. Furthermore, a method for incorporating this uncertainty into the data mining model is required. This is discussed in Sections 7.2 and 7.3.1 respectively.

7.2 Search system

As explained in Section 3.4 and Section 5.3 of Chapter 3 and 5 respectively, a common type of uncertainty related to online data is reference uncertainty. Reference uncertainty can arise when searching for online information, e.g., when multiple similar results are found. To test the effect on prediction accuracy of incorporating this uncertainty into the predictions, a search system that determines the degree of (un)certainty of each result is required. As explained in Section 3.1, H. Been built a prototype of a system that uses data about a real world person to automatically find online manifestations of that person. Figure 7.2 illustrates the general setup of their search system.

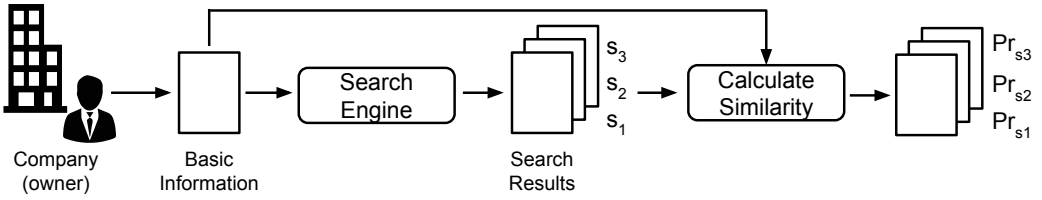

FIGURE 7.2: Search system for determining reference uncertainty.

Basic information known about subjects (e.g., name, address etc.) is used to construct search queries that are passed to the search engine. This search engine will possibly return multiple search results. The information within these search results or the information on the pages to which the search results are referring to are used to determine for each search result the confidence that it is about the subject that was searched for. A similarity score is calculated for each search result based on the basic information known about the subject. The two elements necessary for the search system of Figure 7.2 are discussed in the subsections below.

7.2.1 Search strategy

A search strategy is necessary to find candidate results. The idea of accounting for uncertainty is that, although the correct result is not ranked first, its data is still incorporated in the calculations with some (small) probability. Hence, it is mainly important to get a high recall, i.e., if a person has an account it should be within the search results. As explained in Section 3.1, H. Been built a prototype of a system that uses data about a real world person to automatically find online manifestations of that person. In [13], H. Been describes his search strategy to find online manifestations of people on Twitter using the Google search engine with which he was able to achieve a recall of 91% (i.e., for 91% of the subjects the correct Twitter account was found in the result set and thus included as possible match). He showed that leaving out the first name of a subject reduces recall significantly, which explains why in Section 5.6.1 we were not able to manually find Twitter and Facebook accounts of persons using our data set. In addition, he shows that leaving out the (e-mail) address and telephone number has no effect on recall.

The search queries constructed by H. Been were of the form 'X Y site:twitter.com' and were constructed for every combination of first name, last name and tussenvoegsel[1]. A test was performed to determine whether a similar search strategy also works well with our data set. In contrast to H. Been, the nicknames of persons is not in our data set. The legal first name of company owners that are in our data set often deviates from the person's nickname, which is used on social media. Including these deviating legal first names increases the probability that the correct account is excluded from the search result. Therefore, in this test, the legal first names were excluded from the search queries. However, for LinkedIn, we instead used the company name in the query. When searching for corporate accounts on Twitter and Facebook, we only used the company name for constructing the query. Table 7.1 shows the search strategies that were used in

[1] A tussenvoegsel is a Dutch phenomenon in which a word / a few words are placed between the first and last name.

the test to find a person's/company's account on LinkedIn, Twitter and Facebook and summarizes the results.

TABLE 7.1: Search strategies.

Source	Query	# Subjects	Recall
LinkedIn	last_name company_name site:linkedin.com	91	72.53%
Twitter	company_name site:twitter.com	67	43.94%
Facebook	company_name site:facebook.com	86	63.95%

The table shows the number of subjects that were searched using these queries and the recall of this query strategy. In this test, the top 100 results of these queries in the Google search engine were considered. Legal form abbreviations (such as 'B.V.') were removed from the company name and person names were stripped from accents in this test. Compared to the results of H. Been, the results of Table 7.1 are relatively low. However, in contrast to the work of H. Been, in this test only a single query was sent for each entity. Recall could potentially be improved by sending additional queries for possible subsets of the company name, because in general the legal company name tends to be longer than the common company name.

7.2.2 Similarity matching

H. Been describes several attributes of subjects that can be used for calculating similarity scores of search results, e.g., name, city, language and email. Appendix A explains several functions that are used by H. Been or can be used for calculating the degree of similarity. The information that can actually be used for matching search results depends on the information that is known about the subject, i.e., the information that is available in the data set. For example, in contrast to H. Been, our data set does not contain information on the address, telephone number, e-mail address or aliases of persons that can be used for matching. However, our data set does contain extra information about the company a person is working, e.g., company name, Chamber of Commerce registration number (KVK), street, zip code, city, country, website and sector. This is ideal for matching results of LinkedIn, which is used in this research as information source for persons.

7.3 Uncertainty in data mining

Uncertain data can be stored, queried and combined using the concepts and techniques described in Section 3.4 of Chapter 3. However, by default the classifiers which were used in the experiments of Chapter 6 cannot directly handle uncertain data. These classifiers

assume that data are exact and require instances to be single values. However, in our case we have a set of possibly multiple tuples consisting of a value with a particular probability. In the next subsection, we will discuss two approaches of dealing with uncertainty in data mining. Then we will discuss related work on handling uncertainty in data mining.

7.3.1 Approaches for handling uncertainty

There are two approaches for handling uncertainty in data mining:

1. Transform the data itself before passing it to the classifier. According to Qin et al.[56] and Ren et al. [57], a straightforward method for dealing with uncertainty is to replace the uncertain data with its expected value. Because this is a single number, traditional classifiers can be used. Although this is a relatively simple method of dealing with uncertainty, it might also result in valuable information loss.

2. Change the classifier itself so that it can handle uncertain data. Some available related work on classifiers adapted for uncertain data is discussed in Section 7.3.2. Although this approach might give more accurate results, implementations of the listed uncertainty adapted classifiers are not freely available.

The features used in this research were discretized into bins. Figure 7.3 gives an overview of both aforementioned approaches for handling uncertainty in data mining when combined with a discretization step.

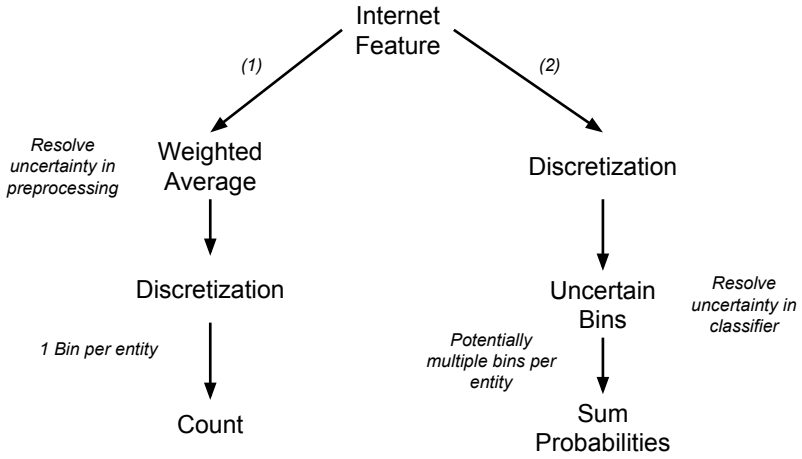

FIGURE 7.3: Approaches to handle uncertainty in data mining.

Whenever the discretization is applied after calculating the expected value, it might have an impact on the accuracy of the classifier. For example, assume that the four training

instances of an attribute belonging to a particular prediction class as depicted in Table 7.2.

TABLE 7.2: Handling uncertainty in case of discretization.

Instance#	**Uncertain Indicator Value**	**Uncertain Bins**	**Weighted Average**	**Weighted Average Bin**
1	(5, 0.6) (14, 0.3) (18, 0.1)	[5,10) [10,15) [15,20)	9	[5,10)
2	(14, 0.8) (7, 0.2)	[10,15) [5,10)	12.6	[10,15)
3	(0, 0.51) (10, 0.49)	[0,5) [10,15)	4.9	[0,5)
4	(19, 0.6) (13, 0.4)	[15,20) [10,15)	16.6	[15,20)

Assume that each of those instances has a number of possible values with a certain probability (depicted by tuples in the table). Furthermore, assume that the data is discretized into four equal-width bins. In case the classifier can handle uncertain data, it will take each bin derived from discretizing the value into consideration with its corresponding probability. This is illustrated in part (a) of Figure 7.4. As can be seen, the bin $[10, 15)$ is by far the most likely bin for that particular rating class. However, if we would first calculate the weighted average of each instance and use that value to determine the bin, we get a different image. This is illustrated in part (b) of Figure 7.4. In this case, all bins seem equally likely and therefore the classifier cannot identify a bin that is a predictor for the particular prediction class. Note that this example shows an extreme case and that in practice the weighted average approach might already show how accounting for uncertainty affect prediction accuracy.

7.3.2 Related work on resolving uncertainty in classifiers

This subsection discusses related work on variants of classifiers used within this research that can handle uncertain data.

Decision tree classifier

In [58], Qin et al. present a Decision Tree for Uncertain Data (DTU), which is an extension of the C4.5 algorithm that allows it to handle both uncertain numerical attributes

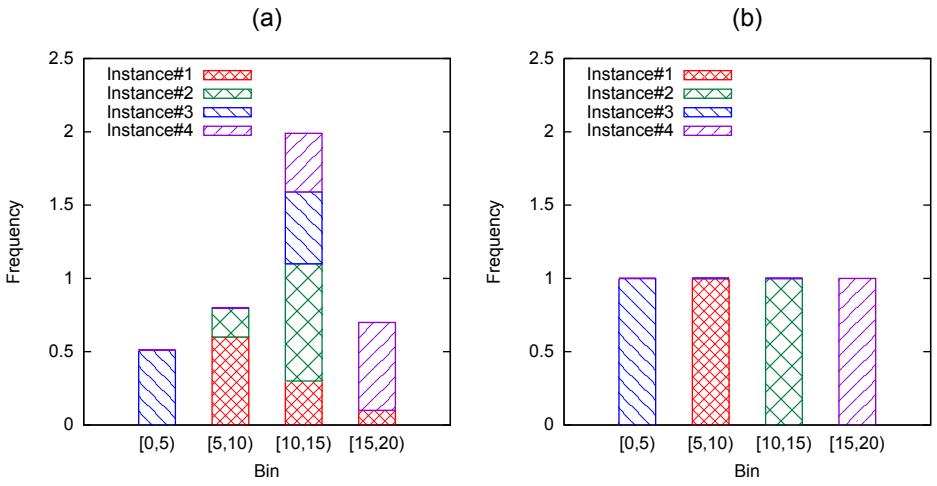

FIGURE 7.4: Handling uncertainty in case of discretization.

and uncertain categorical attributes. One of the main aspects of a decision tree is to select at each tree-grow step an attribute and determine how to split the records using some splitting measure, for example, information entropy and the Gini index. These default splitting measures cannot handle uncertain data. Qin et al. solve this by introducing probabilistic cardinality and use this to compute probabilistic entropy. The probabilistic cardinality of a data set of a particular partition is the sum of probabilities of each instance whose corresponding uncertain numerical / categorical attribute falls into that partition. In addition to the algorithm for tree construction also the prediction process needs to be adapted. As commonly, this process starts from the root node. An appropriate path is chosen based on the outcome of the test condition that is applied at each node. Whenever the attribute that is tested by a test condition is certain, choosing a branch is straightforward. However, when the attribute is uncertain and the branch to be chosen is ambiguous, both branches are chosen with their corresponding probability. Hence, you can end up in multiple leaves. For each leaf node, the probability is calculated of ending up in that node by taking into account the probability of nodes along the path. The probability of each output class is determined by summarizing the probability of ending up in corresponding leaf nodes. The class with the highest probability will be predicted.

Rule-based classifier

In [56, 59], Qin et al. present uRule, which is based on the commonly used RIPPER rule-based algorithm. The growing and pruning measures for the rule learning procedure of the RIPPER algorithm were extended such that it can handle both uncertain

numerical attributes and uncertain categorical attributes. Their algorithm extracts rules one prediction class at a time for a data set. It starts with an empty rule for every prediction class. Then, as with DTU, the uRule algorithm uses probabilistic cardinality to determine the information gain of an attribute. It determines, based on this information gain, the attribute and split point to be added into the rule condition. In contrast to traditional rule-based algorithms where instances are either fully covered or not by a rule, in uRule a tuple can be partially covered. A special function was defined to handle this partial coverage of uncertain data by a rule. The part of the instance that is (partly) covered by the rule is removed from the data set, and the rule growing process continues, until either all the data are covered or all the attributes have been used as condition. In addition to the construction phase, also the prediction phase of the classifier was adapted such that it can handle uncertain data. In case of partial coverage of instances by rules and in contrast to normal rule-based classifiers, uRule might trigger multiple relevant rules to compute the probability (using probabilistic cardinality) for the instance to be in each class. It then predicts the instance to be the class with the highest probability.

Naive Bayes classifier

Also variants of the Naive Bayes classifier that can handle uncertain data have been developed by both Ren et al. [57] and Qin et al. [60]. Both proposed methods to handle uncertain numerical data by using the probability distribution as conditional probability. In [57], Ren et al. tested the performance of their adapted classifier on uncertain data and compared it to the performance of the standard Naive Bayes classifier. In the latter case, they averaged the data so that the standard Naive Bayes classifier can handle the data. They show that including the probability density function of the uncertain data can produce models with higher accuracies. Also the results of Qin et al. suggest that including the probability density function into the Naive Bayes classifier can result potentially in higher accuracies.

Support Vector Machine

In [61], Zhang et al. introduce the total support vector classification (TSVC) algorithm to handle uncertain data in support vector machine classification. The method is named after the total least square method to which it is related. It assumes that inputs are subject to additive noise and that this noise follows a certain distribution. They consider the uncertainty to follow a (bounded) Uniform distribution. This uncertainty is incorporated into the support vector classifier using a specially developed kernel function

(kernel functions define how data is mapped to a higher dimensional feature space such that the separability of the data improves). Experiments show that TSVC performed overall better than standard support vector classifiers.

7.4 Conclusion

Uncertainty can arise when searching for online information. In this chapter we discussed an experimental setup for testing the effect on accuracy of applying the possible worlds concept to the creditworthiness prediction model that uses online data. In this proposed setup, the accuracy of the creditworthiness prediction model is tested both using solely information of the best matching alternative and using information of multiple alternatives together with their corresponding probabilities of being correct. Based on the work of H. Been[13], a search system for finding candidate results and determining the level of (un)certainty of results matching the entity that was searched for is described. In this search system, basic information known about a subject is used to construct search queries and to determine a matching score. Resolving uncertainty before passing data to the classifier and resolving uncertainty in the classifier itself were discussed as approaches for dealing with uncertainty in the experimental setup. Testing how uncertainty affects the prediction accuracy using the experimental setup described in this chapter depends on the data mining models constructed in the previous chapter. In order for conclusions about the effect of uncertainty on accuracy to be relevant, some baseline performance is required. Because none of the models constructed in Chapter 6 performed better as the baseline, we were not able to test how uncertainty affects the prediction accuracy of our creditworthiness prediction models.

Chapter 8

Ethical Considerations

Using online information for making creditworthiness predictions can have several ethical concerns. Two of these concerns are discrimination and privacy violations, which are discussed in this chapter.

8.1 Discrimination

Two forms of discrimination can be identified [62]:

- Disparate treatment. This is intentional discrimination in which practices or policies would cause two similarly situated people, of which only one is a member of a particular protected class (e.g., race, gender, ethnicity, religion, age), to suffer a different fate [62, 63].
- Disparate impact. This is unintentional discrimination. Practices or policies that appear to be neutral are discriminatory in its application or effect and causes people to be treated differently based on their membership to a protected class [64, 65]. In case of disparate impact, the effect of the policies/practices have a disproportionately adverse impact on protected classes[66]. The Equal Employment Opportunity Commission in the United States has defined a "fourth-fifth rule" for determining cases of disproportionately adverse impact in the employment selection procedure [67]. According to this rule, there is adverse impact whenever the selection rate for a particular protected class is less than 80% of the group with the highest rate. However, detecting under- and over-representation of members of protected classes is not always evident [62].

The distinction between disparate treatment and disparate impact might not always be obvious. Data mining could be used to mask intentional discrimination, such that it is undetectable due to the complex models build by data mining tools, or at least defensible (e.g., data is objective [68] and algorithms are neutral [69]) [70]. Instead of using the membership to the protected class as input to a model, a proxy feature (i.e., feature that by itself is no protected class, but can indirectly be used to distinguish between members and non-members of a protected class) could be used to mask intentional discrimination. Discrimination by persons is then replaced with discrimination by computers. Tene et al. [71] note that discrimination is not necessarily always undesirable. In some cases, discrimination could also be socially desired (e.g., discount for children and elderly) or generally acceptable (e.g., personalized recommendation systems).

The models in this research were not created with the intention to discriminate between protected classes, hence there is no disparate treatment. However, the models created could result in a disparate impact. To solve a problem, data miners translate the problem into a question about the value of a target variable [62]. Decisions regarding this target variable can result in unintentional discrimination. For example, the method in which PDs are translated into rating classes that have a particular meaning could result in a particular class of companies to be less often creditworthy. In this research, we used an existing rating model that is used in practice and for which there are no indications that it suffers from these issues.

In addition, the Internet features selected in our models could result in disparate impact. For example, younger people most often have less months of working experience than older people. Hence, incorporating this feature could be a proxy for age and result in certain age classes being discriminated. We could not test whether this feature has a disparate impact, because our data set does not contain the age of the persons. However, when such a prediction system would be used in practice, from an ethical point of view it is advised to first test whether the model does not result in disparate impact, e.g., using the guideline of the Equal Employment Opportunity Commission.

The training data itself can also result in discriminatory models. If the training data is a biased sample of the intended population for which it will be used, the predictions may be disadvantageous for those who are under- or over-represented in the data set[62]. For this reason, the filters that were applied in Section 4.3.2 to increase the quality of the data set were chosen carefully. Most filters that were applied were to select only companies for which the model is intended. Other filters (e.g., excluding companies for which the balance sheet is unbalanced) were used to further improve the quality of the data set and concerned various companies instead of only a few types of company.

8.2 Privacy

Collecting data from social media can result in several privacy violations. Frequent causes for these violations are: collecting too much information; collecting information in an unethical manner; using information for purposes other than have been indicated when collecting the data; and not allowing subjects to correct errors in the information that was gathered [72].

Wynsberghe et al. [72] discusses ethical limits for using data obtained from social media and proposes guidelines for incorporating ethics into research involving data from social media. The guidelines they proposed for best practice when using data from social media can be translated into five questions:

1. What are the key actors? (direct and indirect subjects, researchers, etc.)
2. What is the context and what does privacy mean in this context? (location and data content)
3. What is the type and method of data collection? (passive vs active)
4. What is the intended use of the info and the amount of info collected?
5. What are the intended values? (making explicit and scrutinizing intended values of the researchers)

1. Key actors

The key direct actors, besides the researcher himself and Topicus Finance, are the credit lender and company/entrepreneurs applying for a loan. Indirect actors are the persons related to the company, but who are not involved in the credit application (e.g., employees of the company). Searching for online manifestations might also yield incorrect search results. These unintended subjects that are collected while searching for information are also indirect actors. Furthermore, although the companies / entrepreneurs of our data set are no indirect actors when the system is used in practice, for this research they are because they did not apply for a loan and hence are not directly involved in the system.

2. Context

The contexts in which the system is working are LinkedIn, Twitter and Facebook. LinkedIn is a social media platform which as a goal has to connect the world's professionals [1]. For default LinkedIn members, the profile visibility is based on the connection degree and often is more limited than that for premium LinkedIn members. According to their privacy policy, LinkedIn's premium service allows enterprises and professional organizations to view profile information and store information they have independently obtained about you outside of their services. Although we gathered the information from LinkedIn through a proxy, it in general is accessible by LinkedIn partners. Hence, LinkedIn users should not expect that their profile is protected against companies.

Twitter is about sharing content with the world [2]. Therefore, users should not have the impression that information is only accessible to a particular set of individuals.

Facebook mainly is about staying connected with friends and family[3]. In this research, we studied the use of company Facebook pages. These company pages are about sharing content about the company with the world. Hence, companies should not have the expectation that the information is protected.

3. Type/method data collection

The type of data collection in this research is passive. The companies and persons related to the company for which information was harvested were not notified. The data collected in this research is used for training and testing general models and not already to make a prediction about individual companies. Furthermore, the Dutch privacy law (Wet Bescherming Persoonsgegevens)[73] concerns personal data relating to natural persons only. Therefore, the information of corporate Twitter and Facebook pages cannot be considered personal data. In addition, the context of LinkedIn, from which personal information is collected, is about coming into contact with companies. Hence, it was found justifiable to collect this data passively for the experiments in this research.

4. Intended use

The information collected during the experiments is used for training and testing our creditworthiness prediction model. It is stored for a limited period of time during the

[1] `https://www.linkedin.com/legal/privacy-policy`
[2] `https://twitter.com/tos`
[3] `https://www.facebook.com/facebook/info?tab=page_info`

research. The intended use for the researcher and Topicus Finance is to test the added value of applying the possible worlds concept of probabilistic databases within the financial domain. The intended use of the credit lender for the online data is to improve the existing risk estimation model or use it as an alternative for it. The information is intended to be used to obtain insight on the creditworthiness of the company. Even the information gathered about persons, is gathered about persons in the role as an entrepreneur to gather insight on the creditworthiness of the company. Furthermore, the indicators currently extracted from the information sources are mostly numerical values and cannot be considered personal data as in the definition of the Dutch privacy law (Wet Bescherming Persoonsgegevens): "any information relating to an identified or identifiable natural person"[73]. Individual persons cannot be identified from these numerical values, especially considering the discretization step that is applied before training and testing the model (see Section 6.1.1). Refer to Section 5.7 for the specific data collected from LinkedIn, Twitter and Facebook.

5. Intended values

The intentions of the researchers of gathering the online data is to construct a creditworthiness prediction model such that the impact of modeling uncertainty can be tested. The intended value is to improve the reliability (i.e., a more accurate reflection of the actual creditworthiness) of creditworthiness predictions of companies by building a prediction model that uses uncertain online data. Given these intentions, the policy that LinkedIn information can be used by companies, the fact that for Twitter and Facebook only corporate accounts are used and the fact that the extracted indicators are discretized numerical features, we concluded that it was ethically justifiable to collect and use the Internet features of subjects in our data set.

Also when such a system is used in practice with, for example, other information sources or Internet features, we think it can be justifiable. The intended value for entrepreneurs applying for a loan is a fairer prediction system such that entrepreneurs pay a fairer amount of interest. Companies that according to the online data are more creditworthy than based on the financial figures alone can be asked to pay less interest, while companies who are found to be less creditworthy based on the online data can be asked to pay more interest. This increases the level of justice, because creditworthy companies have to contribute less to covering the risks of the credit lenders of lending money to non-creditworthy companies.

People related to the company of which online information is collected may experience it as a reduction of their privacy. However, these people also have an interest in that it goes

well with the company so they can stay employed. As explained in Chapter 7, searching online information is accompanied with uncertainty. The idea behind this research is to use the uncertain candidate results in the creditworthiness prediction model. By accounting for uncertainty, justice increases because more often the information of the correct subject is included into the prediction. The indirect subjects who are not (or no longer) related to the company but on which data is collected while searching for information may experience it as a reduction in their privacy, because their information could be used in the judgment of someone else's creditworthiness. However, even for persons that are not related to the company the improved prediction model can have a positive effect, because more creditworthy companies can get a loan which allows them to invest and expand which is good for the economy as a whole. Furthermore, an improved prediction model prevents more non-creditworthy companies from obtaining a loan and as a consequence from making more debts, e.g. with suppliers. So, it limits the consequential damage of a bankruptcy of the company which again is also good for the economy as a whole.

Although this intended use was not explicitly discussed, an online data prediction model can also have advantages when used as an alternative to existing creditworthiness prediction models instead of an addition. For example, due to the large number of applicants for a loan (e.g., in developing countries), using traditional creditworthiness prediction models might not always be a viable option for timely processing credit applications [16]. In that case, companies / entrepreneurs who would be entitled to obtain a loan are not able to obtain it. Using online information can be used as a faster alternative to the standard creditworthiness prediction models, such that more creditworthy companies / entrepreneurs can get a loan which in turn can contribute to the economy of a country.

Chapter 9

Conclusions

This research is about using uncertain online information as a predictor of a company's creditworthiness and determining whether *Given a limited and uncertain set of online data about a real world person or company, can a prediction of their creditworthiness be made using this data?* Four subquestions were defined to answer this main research question. The main research findings are discussed in the following subsection.

9.1 Main research findings

Q1: Which financial data related to creditworthiness is available to test the performance of online information as predictor for a company's creditworthiness? Credit ratings were found to be a possible target variable of creditworthiness prediction models and are used in this research for training and testing models that use online information for predicting a company's creditworthiness. To obtain these ratings, we imported balance sheet and profit & loss account figures from an external data provider for a set of Dutch companies. Ratings were calculated over these financial figures using rating formula to which Topicus Finance has access. A data set consisting of well-distributed ratings of 3579 different companies was constructed and is available for testing the prediction performance. Based on related work [16], it was concluded that for data mining the size of this set should be sufficient.

Q2: Which online available information can be used as a predictor for the creditworthiness of these companies? A survey was held among 17 employees of Topicus Finance and literature was consulted to identify possible information sources. The results of this survey and literature study suggest that social media is an interesting source to use for creditworthiness predictions. It was explained that in order to train

and test the data mining model, it is required that information within an information source is available for a significant amount of companies from our financial data set. Hence, a test was performed using 50 random companies from our data set to determine the presence and information availability for the identified information sources. LinkedIn accounts of persons related to a company from our sample were found in 84%. Furthermore, for 44% and 45% of the companies in our sample a corporate Twitter and Facebook account was found respectively. From these results, we concluded that personal LinkedIn profiles and corporate Twitter and Facebook accounts can be used for constructing a prediction model. A data set consisting of the indicators from these sources for 387 different companies and 436 persons related to these companies was constructed by manually searching their social media accounts, harvesting the information and extracting the indicators from these information sources. Based on literature and the information availability within personal LinkedIn accounts, we decided to extract as an indicator: the number of connections, skills, endorsements, months experience, job switches, education switches; and the education level and the presence of a summary. In particular numerical indicators expressing the volume of posts, influence and engagement were extracted for corporate Twitter and Facebook social media accounts.

Q3: What is the performance of these indicators as a predictor for creditworthiness? Two experiments were performed to answer this question. In the first experiment, solely the selected Internet features were used to build a prediction model using the Naive Bayes, J48, Random Forest and SVM classifiers. The accuracy of these models was determined by calculating the fraction of correctly classified instances. These models had an accuracy of 21%, 27%, 26% and 31% respectively. Their accuracy was compared to a ZeroR baseline model, which always outputs the rating class it found to be most occurring. The results showed that none of the models achieved a higher prediction accuracy as the 31% of the ZeroR constructed model.

In the second experiment, the same classifiers were used for testing whether these Internet features might have predictive value in combination with financial data. Prediction models were created based on solely financial data and on financial data in combination with selected Internet features. The performance of both models was compared by determining their approximation accuracy to an existing rating model. The two best performing financial models were built using the Random Forest and J48 classifiers and had an approximation accuracy of 68% and 63% respectively. Adding Internet features to these models gave mixed results with a significant decrease and an insignificant increase in approximation accuracy.

From the results of both experiments, we concluded that using the selected Internet features, either in combination with or without extra financial company figures, does

not have an added value for predicting a company's creditworthiness in this setup.

Q4: How does incorporating the uncertainty of online data affect the prediction accuracy? An experimental setup for testing the effect on accuracy of applying the possible worlds concept to the creditworthiness prediction model that uses online data was described. In this setup, the accuracy of the creditworthiness prediction model is tested both using solely information of the best matching alternative and using information of multiple alternatives together with their corresponding probabilities of being correct. A search system, which is based on the work of H. Been [13], was described as a method for finding candidate results and determining the level of (un)certainty of results matching the entity that was searched for. Two approaches were discussed to deal with uncertainty in data mining: resolving uncertainty while preprocessing the data or resolving it in the classifier itself. Testing how uncertainty affects the prediction accuracy using the experimental setup described in this chapter depends on the data mining models constructed in the previous chapter. In order for conclusions about the effect of uncertainty on accuracy to be relevant, some baseline performance is required. Because none of the models constructed in Chapter 6 performed better as the baseline, we were not able to test how uncertainty affects the prediction accuracy of our creditworthiness prediction models.

The conclusion to the main research question is: given the setup using ratings and our selection of online data about a real-world person or company, we were not able to make an accurate prediction of a company's creditworthiness. However, in the next subsection we propose several research directions that we believe might allow a prediction to be made.

9.2 Discussion and Future work

The results of this research showed that the selected Internet features cannot be used to make a prediction of a company's creditworthiness using our setup. However, this does not imply that these features cannot have a predictive value in other setups or using other and larger data sets.

Based on the results of Heijnen that showed that companies of various industries are active on social media [4], we assumed that companies of various industries and also various sizes could be used to create a single prediction model for all these types of companies. However, this might not be the case. When manually studying the relation between the probability of default and indicator values of the attributes selected by the attribute evaluator of Section 6.1.1, we noticed that for all features the indicator values

seem to be scattered over both low and high probabilities of default. The scattering of these values is shown in Figures F.1, F.2 and F.3 in Appendix F. However, the attribute evaluator did select several features, so apparently these have some added value. In addition, when studying the J48 decision tree of the financial model of Section 6.2.2 which includes the selected Internet feature, we found that the selected Internet feature (#Months experience) is located at depth 2 of this tree with a maximum depth of 7. In a J48 decision tree, attributes are ranked according to their predictive value. Nodes closer to the root of the tree are found to be more predictive than nodes more downward in the tree. This means that this Internet feature was found to be more predictive than several financial features.

As explained in Chapter 7, we were not able to test the effect of uncertainty on prediction accuracy because the models constructed Chapter 6 did not perform better as the baseline. Using Figure 6.4 in Chapter 6 we showed that it is not likely that this is caused by a fluctuation in ratings. However, there are several research directions that could be explored to build a more accurate model for predicting a company's creditworthiness and which might allow the effect of accounting for uncertainty to be tested.

Sample size

As was shown in Section 6.1.1, it turned out that there were too few samples for each rating class to accurately train the classifier for each class. It was shown that the prediction models of all tested classifiers that solely used Internet features did not perform better than the model of the ZeroR classifier, which always outputs the most-frequent rating class. It is expected that when more training samples are available for the other rating classes, that the classifiers will be better able to predict those classes and possibly perform better than the ZeroR classifier. As explained in Section 4.3.2, in this research all AAA and D ratings were left out of the data set. Among those left out companies are also companies that do not suffer from the issues mentioned in that section (e.g., on paper are no holdings but in practice are). To increase the number of samples for the outer classes, one could selectively leave in these ratings. However, an objective filtering criterion has to be defined in order to not produce biased results.

Information on more entities / Predictions on smaller companies

In this research we tried to relate the extracted Internet features of a few (important) company related persons to the company rating. Based on literature described in Section 5.2.2 and Section 3.2, social media information about persons seem to be eligible as creditworthiness predictor for persons. The reason why we were not able to use this

social media data to build an accurate model could perhaps be explained by the fact that a company is a larger entity than a person. Hence, for many companies the creditworthiness relies on more than only the creditworthiness of a few people. We suggest that the prediction accuracy might be improved by using either the information of more people related to the company, for example, all employees, or focusing on the creditworthiness of freelancers (Zelfstandige Zonder Personeel). With the data set available in this research, we were not able to find all employees of a company.

More sophisticated Internet features

In this research, we attempted to make a creditworthiness prediction based on relatively simple (mostly numerical) Internet features, because these could most easily be extracted. It turned out that with our data set and setup these simple features could not be used to make an accurate prediction of a company's creditworthiness. However, although it was out of scope for this research, also other more complex features could be extracted from these same sources by performing a more detailed analysis. For example, sentiment analysis in posts/tweets or other forms of text analysis and natural language processing (e.g., searching for specific events or subjects) might prove to be much better predictors. Once a reasonable model is built, a higher accuracy could potentially be achieved by additionally incorporating certain online data (e.g., prejudgments about people and companies or using inflation and employment statistics).

Personal Twitter and Facebook profiles

Our data set did not contain enough information to find the Twitter and Facebook account of persons related to the company. This was partly caused by only knowing the legal first name. In addition, whenever multiple results of persons with the same name exist, we do not have any other information about the person to determine which of these results is the correct match. Because according to related work these personal profiles also contain valuable information for predicting creditworthiness, it might be worth researching features extracted from these sources whenever the available data allows for testing this.

Prediction model per industry sector

In addition to the PD which can vary per industry sector and for which we checked in Section 5.7 that it does not vary too much per sector, also the values of Internet features themselves can vary per sector. While manually constructing the data set for training

and testing the data mining model in Section 5.7 we noticed that the value of these Internet features varies widely per industry. For example, most construction companies which do have a social media account seem to be far less active than, for example, wholesale & retail trade companies who are active on social media. Furthermore, also the value of Internet features seems to differ per company size. For example, for larger (inter)national companies, it is common to have thousands of likes, while for smaller local companies this is often only a few hundred. Although these values seem to differ per sector/company size, we can have creditworthy and non-creditworthy companies for all sectors/company sizes. This means that for some sectors/company sizes a certain value might indicate that the company is creditworthy while for other sectors/companies of other sizes it might not. We suspect that a better accuracy can be achieved when constructing a model for a more specific set of companies. In this research we did not focus on a specific sector of companies, because both the amount of companies for which a rating is available was limited and we wanted to hold an unbiased view.

Appendix A

Similarity Functions

This appendix discusses several similarity functions for matching search results of the search system of Section 7.2 in Chapter 7 to the subject that was searched for.

A.1 N-grams

N-grams are sequences of adjacent items (i.e., sub-strings) of length n constructed from a sequence of text [74]. These items can for example be letters or words. Well-known n-grams are unigrams (n = 1), bigrams (n = 2) and trigrams (n = 3). For example, possible bigrams for the sequence of letters 'janssen' are: 'ja', 'an', 'ns', 'ss', 'se' and 'en'. Determining the similarity between two strings based on n-grams is done by calculating some similarity measure over the number of n-grams. A commonly used similarity measure is the Jaccard index, which takes the intersections of two sets and divides it by the maximum length of both these sets:

$$sim_{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{A.1}$$

In this case, it is the number of n-grams the two strings have in common divided by the number of n-grams of the larger of the two strings. To calculate the n-gram similarity of, for example, the string 'janssen' with the string 'jansen', you calculate the possible bigrams of both strings. The possible bigrams for the string 'janssen' were shown above. The possible bigrams for the string 'jansen' are: 'ja', 'an', 'ns', 'se' and 'en'. These strings have 5 bigrams in common. The number of bigrams of the largest string is 6. So the similarity of these strings according to this measure is $\frac{5}{6} = 0.83$

A.2 Levenshtein distance

The Levenshtein distance is a metric that calculates the smallest number of edit operations required to transform one string into another [74, 75]. Edit operations for this metric are insertions, substitutions and deletions of single characters. Given two strings s_1 and s_2 and let $dist_{Levenshtein}(s_1, s_2)$ be the Levenshtein distance between these strings. A similarity measure can be calculated using the following formula:

$$sim_{Levenshtein}(s_1, s_2) = 1.0 - \frac{dist_{Levenshtein}(s_1, s_2)}{max(|s_1|, |s_2|)} \tag{A.2}$$

For example, let $s_1 = Mark\ Rutte$ and let $s_2 = M@rkRutte67$. One substitution is necessary to transform the '@' into an 'a', two deletions are necessary to remove '67', and one insertion is necessary to add the space when transforming s_2 into s_1. Thus, the Levenshtein distance equals $dist(s_1, s_2) = 4$. Furthermore, $max(|s_1|, |s_2|) = max(10, 11) = 11$. So, the similarity measure equals $sim(s_1, s_2) = 1.0 - \frac{4}{11} \approx 0.64$.

A disadvantage of the Levenshtein distance is that it does not perform well when entire segments of a string differ [76], for example, due to the use of abbreviations (*Pieter Jan Bos* vs. *Pieter J Bos*) and prefixes (*MSc. Jan Bos* vs. *Jan Bos*). This is due to the equal weight the Levenshtein distance assigns to the edit operations and to the fact that each character is considered individually.

The Damerau–Levenshtein distance is a variation of the Levenshtein distance in which the transposition of two adjacent characters is also considered as an operation [77]. For example, two edit operations are necessary when calculating the similarity between *Makr* and *Mark* using the default Levenshtein distance. However, only one edit operation is necessary to swap 'k' and 'r' in *Makr* when calculating the similarity using the Damerau-Levenshtein distance. This method can for example be useful in fraud detection.

A.3 Jaro–Winkler distance

The Jaro distance [74, 76, 78] is a similarity metric that first identifies the number of common characters in two strings. Then, it determines the number of transpositions necessary to rearrange these common characters such that they are in the same order as in one of these strings. Characters are considered common when they are equal and are within half the length of the shorter string [76].

More formally, let s_1 and s_2 be two strings. Let c be the set of common characters that are within half the length of the longer string. Furthermore, let t be the number of

transpositions. The Jaro similarity measure is calculated as follows:

$$sim_{Jaro}(s_1, s_2) = \frac{1}{3} \times (\frac{|c|}{|s_1|} + \frac{|c|}{|s_2|} + \frac{|c| - 0.5t}{|c|}) \tag{A.3}$$

For example, let s_1 = *Dr. Jan Bos* and let s_2 = *Prof. Jan Bos*. Then, $|s_1| = 11$ and $|s_2| = 13$. So, the maximum distance between equal characters for them to be considered common is $\lfloor 0.5 \times min(11, 13) \rfloor = 5$. All equal characters in this example are within this 5 character range, hence $c = \{r, ., J, a, n, , B, o, s\}$ and $|c| = 10$. These common characters are in the same order in both strings, so $t = 0$. The Jaro similarity is: $sim_{Jaro}(s_1, s_2) = \frac{1}{3}(\frac{10}{11} + \frac{10}{13} + \frac{10-0}{10}) \approx 0.89$. However, the Levenshtein similarity of these strings is only $sim_{Levenshtein}(s_1, s_2) = 1.0 - \frac{3}{13} \approx 0.77$.

The Jaro distance is a commonly used similarity measure for name matching [74, 78]. However, it does not perform well for longer strings separating common characters due to the distance restriction. [76].

The Jaro-Winkler distance is an improvement of the Jaro distance which also takes into consideration that it is less likely for differences to occur at the beginning of two words than in the rest whenever these words are similar [79]. It does this by assigning more weight to similar initial characters [78, 80]. Let p be the length of the common prefix, up to 4 characters. The Jaro-Winkler similarity measure is calculated as follows:

$$sim_{JaroWinkler}(s_1, s_2) = sim_{Jaro}(s_1, s_2) + \frac{p}{10} \times (1.0 - sim_{Jaro}(s_1, s_2)) \tag{A.4}$$

For example, let s_1 = *Jan* and let s_2 = *Jan Bos*. Then, $sim_{Jaro}(s_1, s_2) \approx 0.81$ and $p = 3$. The Jaro-Winkler similarity score equals: $sim_{JaroWinkler}(s_1, s_2) \approx 0.81 + \frac{3}{10} \times (1.0 - 0.81) = 0.87$

A.4 Term frequency–inverse document frequency

Term-frequency is a metric that computes scores between two documents (or a query term and a document) based on the number of occurrences of this term in the document [76, 81]. The idea behind this metric is that tokens that occur more often are more relevant to a certain context. The order of terms in the document is neglected with this method.

The disadvantage of only considering the term frequency is that all terms are equally important [81]. Terms that occur in most documents are less discriminative and are

therefore less important. The term frequency-inverse document frequency (tf-idf) accounts for this by offsetting the occurrences of words by the frequency they occur in other documents [55, 76, 81].

More formally, let t and d be a query term and document respectively. Let $tf_{t,d}$ be the frequency of term t in a document d. Furthermore, let N be the number of candidate documents and let df_t the number of documents in the collection that contain term t. The inverse-document frequency is defined as:

$$idf_{t,d} = \frac{N}{df_t} \tag{A.5}$$

The tf-idf score combines the term-frequency and inverse document frequency and computed as follows:

$$tf\text{-}idf_{t,d} = \log(tf_{t,d} + 1) \times \log(idf_t) \tag{A.6}$$

The advantage of this method is that it accounts for the distinguishing power of terms [76]. However, it does not perform well when many typographical errors occur in the document.

The cosine similarity measure is often used to transform these tf-idf scores into a single normalized score. Instead of expressing the similarity between a single term and a document, it expresses the similarity between an entire query q and a document d. Two vectors $\overrightarrow{V}(q)$ and $\overrightarrow{V}(d)$ are constructed for this query and document respectively, with one tf-idf calculated score in the vector for each dictionary term in the candidate documents. Let $||\overrightarrow{V}||$ be the length of a vector $\overrightarrow{V}$. The cosine similarity is calculated as follows:

$$sim_{cosine}(q, d) = \frac{\overrightarrow{V}(q) \cdot \overrightarrow{V}(d)}{||\overrightarrow{V}(q)|| \cdot ||\overrightarrow{V}(d)||} \tag{A.7}$$

TABLE A.1: Example of calculating the cosine similarity

(A) Set of candidate documents

DID	Value
d_1	Hilton Hotel
d_2	Golden Tulip Hotel
d_3	InterContinental Hotel
d_4	Ibis
q	Hampshire Hotel

(B) Tf-idf scores for query q and document d_1

TID	Term	$tf\text{-}idf_{TID,q}$	$tf\text{-}idf_{TID,d_1}$
t_1	Hilton	0	0.21
t_2	Hotel	0.03	0.03
t_3	Golden	0	0
t_4	Tulip	0	0
t_5	InterContinental	0	0
t_6	Ibis	0	0
t_7	Hampshire	0.21	0

Table A.1a shows an example of a search query q together with some documents (d_1-d_4) containing the name of a hotel chain. To calculate the cosine similarity of this query q with, for example, document d_1, you first need to determine the tf-idf scores of each term

for q and d_1. Table A.1b shows the tf-idf scores of this example. For example, the tf-idf score of the term *Hotel* in query q was determined by first calculating the term frequency in query q. Because the term *Hotel* occurs only once in q, the term frequency equals: $tf_{t2,q} = 1$. Then the document frequency of this term was determined. Since this term occurs in four of the candidate documents in Table A.1a, the document frequency equals: $df_{t2} = 4$. N consists of all documents of Table A.1a including query q, so $|N| = 5$. The inverse document frequency of this term equals: $idf_{t2} = \frac{|N|}{df_{t2}} = \frac{5}{4}$. The tf-idf score of term t_2 in query q can now be calculated as follows:

$$\begin{aligned} \textit{tf-idf}_{t2,q} &= \log(tf_{t,d} + 1) \times \log(idf_t) \\ &= \log(1+1) \times \log(\frac{5}{4}) \approx 0.03 \end{aligned}$$

The other tf-idf scores in Table A.1b were calculated similarly. The columns $\textit{tf-idf}_{TID,q}$ and $\textit{tf-idf}_{TID,d_1}$ form the vectors $\overrightarrow{V}(q)$ and $\overrightarrow{V}(d_1)$ respectively, so

$$\overrightarrow{V}(q) = \begin{bmatrix} 0 & 0.03 & 0 & 0 & 0 & 0 & 0.21 \end{bmatrix}$$

and

$$\overrightarrow{V}(d_1) = \begin{bmatrix} 0.21 & 0.03 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Now we can calculate the cosine similarity:

$$\begin{aligned} sim_{cosine}(q, d_1) &= \frac{\overrightarrow{V}(q) \cdot \overrightarrow{V}(d)}{||\overrightarrow{V}(q)|| \cdot ||\overrightarrow{V}(d)||} \\ &\approx \frac{0 \cdot 0.21 + 0.03 \cdot 0.03 + ... + 0.21 \cdot 0}{\sqrt{0^2 + 0.03^2 + ... + 0.21^2} \cdot \sqrt{0.21^2 + 0.03^2 + ... + 0^2}} \\ &= 0.02 \end{aligned}$$

This similarity score is quite low because the term *Hotel* occurs in most other documents and the query and document do not have other terms in common. However, calculating the Jaccard similarity as in Equation A.1 would result in a much higher similarity score:

$$\begin{aligned} sim_{Jaccard}(\{Hampshire, Hotel\}, \{Hilton, Hotel\}) &= \frac{|\{Hotel\}|}{|\{Hampshire, Hotel, Hilton\}|} \\ &= \frac{1}{3} \\ &\approx 0.33 \end{aligned}$$

Appendix B

Named Entity Recognition and Disambiguation

As explained in Section 3.1 and Section 5.3 of Chapters 3 and 5 respectively, besides uncertainty arising from searching people, uncertainty can also arise from extracting information texts. This appendix describes the three common subtasks in information extraction: Named Entity Recognition (NER), Named Entity Disambiguation (NED) and Fact Extraction (FE)[82].

NER aims to "locate and classify phrases (mentions) in text belonging to predefined categories such as the names of persons, organizations, locations, events, etc." [14, 83]. NED is the task of identifying to which specific person, organization, location, event, etc. is referred to by a mention.

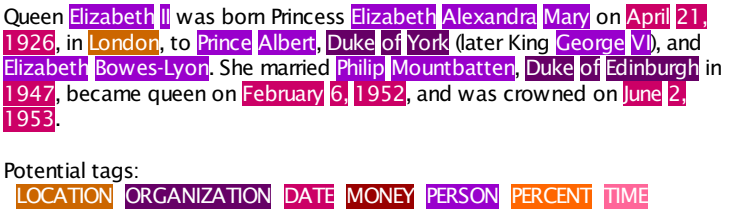

FIGURE B.1: Example of an article about Queen Elizabeth II, tagged with the Stanford NER tool.

Figure B.1 shows an example of an article about Queen Elizabeth II in which the named entities were identified and classified by the Stanford NER tool[1]. Some of the classified entities are ambiguous. For example, the location *London* can refer to the City of London in England or, for example, to London, Kentucky, in the United States. For most humans it is clear that *London* refers to the City of London in this context because

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

we know that Queen Elizabeth II is the Queen of England. However, for a computer this is not that obvious.

One method of recognizing named entities is by defining a set of rules, for example, using regular expressions. This form of named entity recognition only works well for entities with some standardized format such as dates, phone numbers, zip codes and email addresses [14]. Another named entity recognition approach uses machine learning in which the system is first trained to recognize named entities using a training set which is manually annotated. Machine learning techniques include Hidden Markov Models [84], Support Vector machines [85] and Conditional Random Fields [86]. Discussing these techniques in detail is out of scope of this research, because when necessary an existing solution will be used.

Disambiguation of named entities is mostly done using a Knowledge Base (KB) such a Wikipedia or a KB derived from Wikipedia such as DBpedia, Freebase and YAGO [14]. A similarity measure can be defined to compare the context of a mention to the information of an entity candidate in the KB.

Appendix C

Data Mining Classifiers

This section gives some background knowledge about data mining classifiers used within the experiments of Chapter 6.

C.1 Rule-based classifier

Rule-based classifiers generate a set of rules of the form $IF\ condition THEN\ conclusion$ [87]. The rule's condition consists of logically ANDed expressions that each test attribute values. The rule's conclusion contains a prediction class. Rule coverage is defined as the number of instances of a data set that satisfy the condition of a rule. Rule accuracy is defined as the fraction of instances that satisfy the condition and conclusion of a rule. Advantages of rule-based classifiers are that they are relatively easy for people to understand and outperform decision tree learners on many problems [88]. One of the simplest rule-based classifiers is the ZeroR classifier, which is often used as a baseline. The ZeroR classifier ignores all input attributes and always predicts the majority prediction class it found based on training data.

C.2 Decision tree classifier

A decision tree is a tree that can be used to classify data into categories and is a special case of a rule-based classifier, in which each path of the decision tree corresponds to a rule [87]. Most important in decision tree construction is how to identify the attribute that discriminate the instances best (i.e., the attribute with the highest information gain) and what the best criterion to split this particular attribute is. A split criterion at each node of the tree defines the condition of how to divide the data into two or

more parts such that the mixture of prediction classes is minimized. Three commonly used measures for determining splitting conditions are: classification error, Gini index and information entropy. The J48 classifier, which is based on the C4.5 algorithm, uses the entropy measure for constructing decision trees[1]. A common problem with decision trees is over-fitting. Whenever the decision tree algorithm would continue until every leaf node contains only training instances belonging to a particular class, it will generalize to unseen test data poorly. Decision trees are often pruned to reduce this problem. The minimum description length principle (MDL) is an example of an often used pruning technique which defines the cost of a tree as the weighted sum of its error and its complexity.

An advantage of decision trees is that they can be displayed graphically and therefore are easy to explain to people. However, in general, it requires a large amount of training data. Furthermore, according to James et al. [89], it does not have the same level of accuracy as some other classifiers.

The Random Forest classifier tries to overcome this problem by aggregating multiple decision trees. This classifier builds a number of forests using decision trees constructed from bootstrapped training samples. For each tree, instead of all attributes, only a random subset of attributes is chosen as split candidates. The idea behind this is that good predicting attributes are high in all constructed trees of the forest that contain that attribute.

C.3 Naive Bayes classifier

Naive Bayes classifiers follow a simple approach and often outperform more sophisticated classifiers [87, 90, 91]. It is based on Bayes' theorem:

$$P(O|E) = \frac{P(E|O)P(O)}{P(E)} \tag{C.1}$$

$P(O|E)$ is the posterior probability of the outcome given the evidence. $P(E|O)$ is the likelihood of the evidence. $P(O)$ is the prior probability of the outcome. $P(E)$ is the prior probability of the evidence. The numerator of this fraction is of most interest, because the denominator does not depend on the outcome class and hence effectively is constant. The Bayes model is referred to as "naive" because of the assumption of conditional independence [87]. The Naive Bayes classifier calculates the posterior probability $P(O|E)$ using the following formula:

[1] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

$$P(O|E) = P(e_1|o) \times P(e_2|o) \times \cdots \times P(e_n|o) \times P(o) \tag{C.2}$$

The frequency of each evidence attribute per outcome class is determined and from this the evidence likelihood is derived (i.e., calculating per outcome class the frequency of each evidence attribute as a fraction of the number of occurrences of that particular outcome class). The classifier then predicts the class with the highest posterior probability.

Despite the fact that the simplistic assumption that attributes are independent given the outcome class often does not hold, in practice Naive Bayes classifiers perform often quite well especially when combined with attribute selection [90]. Although Naive Bayes can handle numeric attributes, in practice values are often discretized [92]. Missing values are no problem with Naive Bayes because these attributes will be omitted in the calculation. However, performance can deteriorate whenever an evidence attribute does not occur with every outcome class, because it will result in a zero in equation C.2 regardless of the other values.

C.4 Support Vector Machine

Support Vector Machine (SVM) classifiers model instances in a vector space. From these instances, it selects a number of boundary instances of each class and try to find a function that separates them as widely as possible [90]. This is illustrated in Figure C.1. The boundary instances are called support vectors and the boundary is called the

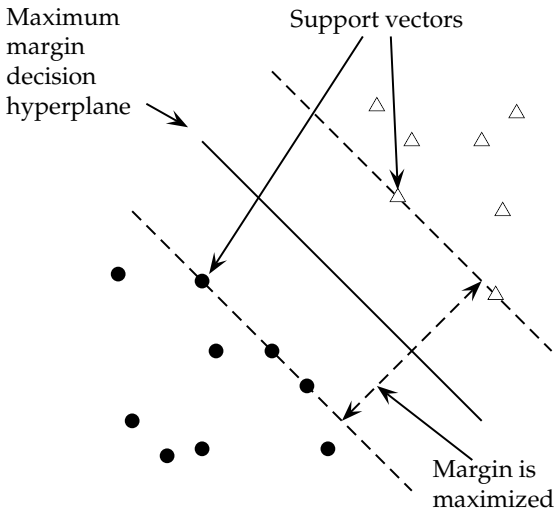

FIGURE C.1: Support Vector Machine.

maximum margin decision hyperplane. It is often hard to find a good boundary between instances without transforming the data first. Therefore, SVMs often map data into a

higher dimensional space such that the classes of data become more readily separable. The kernel function defines how this data is mapped to a higher dimensional feature space such that the separability of the data improves.

Appendix D

Survey Questions

This appendix lists the questions of the survey which was held among employees of Topicus Finance to identify in which online information they are interested in the role as a credit lender. The results of this surveys are discussed in Section 5.1 of Chapter 5 and can be found in Appendix E.

Introduction

Participation to this survey is completely voluntary. Your responses to this survey are confidential and are used anonymously. If you have any questions, you are free to ask them.

Thank you for your time and effort to complete this survey.

Context

Lenders of business loans mainly take the following into consideration:

- The figures supplied by the company
- Information on the entrepreneur/company that is obtained using an interview.

The Internet contains a vast amount of additional information that currently is not taken into account in calculating the creditworthiness of an entrepreneur and his business. The idea of this research is that this information can be used as an alternative and/or addition to existing credit score calculations.

The purpose of this survey is to get an idea of which data sources can be used to estimate an individual's creditworthiness. In other words, what online information could be used by a lender to make an assessment of a person's creditworthiness?

An important part of the problem of this research is to find information in these sources. When searching for this information, it is not 100% certain that results found actually belong to the person or company in question.

I am mainly interested in data sources which got a degree of uncertainty in it. For example, uncertainty that arises when searching for accounts of individuals or companies of which multiple results might actually belong to these persons or companies. Another form of uncertainty can arise in interpreting the text (e.g., does the city name Hengelo refer to Hengelo in Overijssel or Hengelo in Gelderland?).

Your expertise

1. **What is your primary function within the company you work?**

   _____

2. **Do you have experience as a credit lender?** ○ Yes ○ No

3. **Do you have experience in receiving a credit?** ○ Yes ○ No

Information subjects

4. **If you in the role of a lender could choose, about whom would you search information before granting a credit?**

   _____

5. **If you had to choose from the following, about whom would you, in the role of a lender search information before granting a credit? (1 = most important, 2 = least important)**

	1	2	3	4	5	6
The company in question	○	○	○	○	○	○
Other companies within the same sector	○	○	○	○	○	○
Company owner(s)	○	○	○	○	○	○
Family of the company owner(s)	○	○	○	○	○	○
Friends of the company owner(s)	○	○	○	○	○	○
Employees of the company	○	○	○	○	○	○

Information

6a. **Currently lenders mainly use the annual figures of the company and the information exchanged between the loan applicant and lender during an interview. In which online information would you, in the role of lender also be interested?**
I am interested in information about _____

6b. **And why?**
Because, _____

Social media as information source
Facebook

7a. Would you be interested in the role of a lender in the fact that a person or company has a Facebook account?
○ Yes ○ No

7b. If so, why are you interested in it?
Because, _____

8a. Would you, in the role of a lender be interested in information about the size of the Facebook friends network of the person or the company?
○ Yes ○ No

8b. If so, why are you interested in it?
Because, _____

9a. Would you, in the role of a lender be interested in information about the amount of sent and/or received messages of the person or company on Facebook?
○ Yes ○ No

9b. If so, over what period would you look at the amount of activity?
○ The entire period available
○ Only most recent period

9c. If so, why are you interested in it?
Because, _____

10a. Would you, in the role of a lender be interested in the content of messages that a person or business has written and/or received on Facebook?
○ Yes ○ No

10b. If so, for which specific information/topics would you search?
Information about_____

10c. If so, why are you interested in it?
Because, _____

11a. Would you, in the role of a lender be interested in the amount of likes received by a person or company on Facebook?
○ Yes ○ No

11b. If so, why are you interested in it?
Because, _____

12. Is there any other information on Facebook in which you would be interested in the role of a lender?
○ Yes, namely _____
○ No

LinkedIn

13a. Would you, in the role of a lender be interested in the fact that a person or company has a LinkedIn account?

○ Yes ○ No

13b. If so, why are you interested in it?

Because, _____

14a. Would you, in the role of a lender be interested in information about the amount of LinkedIn connections?

○ Yes ○ No

14b. If so, why are you interested in it?

Because, _____

15a. Would you, in the role of a lender be interested in information about the amount of posted messages by the company on LinkedIn?

○ Yes ○ No

15b. If so, over what period would you look at the amount of activity?

○ The entire period available

○ Only most recent period

15c. If so, why are you interested in it?

Because, _____

16a. Would you, in the role of a lender be interested in the education and/or experience listed on the LinkedIn profile of a person?

○ Yes ○ No

16b. If so, why are you interested in it?

Because, _____

17a. Would you, in the role of a lender be interested in the skills that a person or company lists on his profile?

○ Yes ○ No

17b. If so, why are you interested in it?

Because, _____

18. Is there any other information on LinkedIn that you might be interested in the role of a lender?

○ Yes, namely _____

○ No

Twitter

19a. Would you, in the role of a lender be interested in the fact that a person or company has a Twitter account?

○ Yes ○ No

19b. If so, why are you interested in it?

Because, _____

20a. Would you, in the role of a lender be interested in the amount of followers a person or company has on Twitter?

○ Yes ○ No

20b. If so, why are you interested in it?

Because, _____

21a. Would you, in the role of a lender be interested in information about the amount of posted and/or received messages from the person or company?

○ Yes ○ No

21b. If so, over what period would you look at the amount of activity?

○ The entire period available

○ Only most recent period

21c. If so, why are you interested in it?

Because, _____

22a. Would you, in the role of a lender be interested in the content of messages that a person or business has written and/or received on Twitter?

○ Yes ○ No

22b. If so, for what specific information/topics are you searching?

Information about_____

22c. If so, why are you interested in it?

Because, _____

23a. Would you, in the role of a lender be interested in the amount of retweets of messages of a person or company?

○ Yes ○ No

23b. If so, why are you interested in it?

Because, _____

24. Is there any other information on Twitter in which you would be interested in the role of a lender?

○ Yes, namely _____

○ No

Other websites

25a. Are there any other websites where you, in the role of a lender would search for information about the person or the company?

○ Yes, namely _____

○ No

25b. If so, what information are you interested on this/these website(s)?

_____

25c. If so, why are you interested in it?

Because, _____

Appendix E

Survey Results

This appendix lists the results of the survey in Appendix D that were discussed in Section 5.1 of Chapter 5.

TABLE E.1: Survey results of multiple choice questions

Expertise

Nr.	Question	Yes	No
2	Experience as a credit lender?	41.2%	58.8%
3	Experience in receiving a credit?	70.6%	29.4%

Facebook

Nr.	Question	Yes	No
7a	Interested in whether a person or company has a Facebook account?	94.1%	5.9%
8a	Interested in size of Facebook friends network?	52.9%	47.1%
9a	Interested in the amount of posted / received messages on Facebook?	41.2%	58.8%
10a	Interested in content of Facebook messages?	76.5%	23.5%
11a	Interested in the amount of likes received on Facebook?	47.1%	52.9%

Nr.	Question	Entire period	Most recent
9b	Interested in which period of Facebook activity?	41.2%	35.3%

LinkedIn

Nr.	Question	Yes	No
13a	Interested in whether a person or company has a LinkedIn account?	94.1%	5.9%
14a	Interested in the amount of LinkedIn connections?	58.8%	41.2%
15a	Interested in the amount of posted / received messages on LinkedIn?	47.1%	52.9%
16a	Interested in the education / experience listed on LinkedIn?	88.2%	11.8%
17a	Interested in the skills listed on LinkedIn?	70.6%	29.4%

Nr.	Question	Entire period	Most recent
15b	Interested in which period of LinkedIn activity?	35.3%	17.6%

Twitter

Nr.	Question	Yes	No
19a	Interested in whether a person or company has a Twitter account?	82.4%	17.6%
20a	Interested in amount of Twitter followers?	70.6%	29.4%
21a	Interested in amount of posted / received tweets?	41.2%	58.8%
22a	Interested in Tweet content?	64.7%	35.3%
23a	Interested in amount of retweets on Twitter?	58.8%	41.2%
Nr.	**Question**	**Entire period**	**Most recent**
21b	Interested in which period of Twitter activity?	23.5%	23.5%

TABLE E.2: Survey results of open questions

Nr.	Question	Results
1	Function within company	• Project manager. • Financing professional. • Administrative assistant. • Model builder. • Scrum master. • Product owner. • Tester. • Software engineer
4	Additional information subjects	• Suppliers & customers. • Other financiers. • Country. • Past owners.
5	Information subject priority	1. The company in question. 2. Company owner(s). 3. Employees. 4. Other companies within the same sector. 5. Family of the company owner(s). 6. Friends of the company owner(s).
6	Online information	• Moral. • Communication expressions. • Working past / experience. • Living area. • Date of establishment. • Regulations. • Sector outlook. • Presence on (social) media. • Character of the entrepreneur. • Experience of the entrepreneur. • Recommendations received. • Metadata.
	Facebook	
Nr.	**Question**	**Results**
7b	Facebook as information source	• Yes: – It might contain additional information and could perhaps be used for validation of already known information. – Only to check if it does not contain any striking content. • No: It is hardly relevant, because it focuses on private life.

8b	Amount of friends	A larger network might indicate that people are more social, engaged an popular, however, they might also have less time for managing their company because it takes more time to maintain this network.
9c	Amount of posted / received messages	A trend in the amount of messages might say something about the performance of the company.
10c	Message content	• Company performance. • Striking information. • Treatment of employees. • Complaints and compliments. • Building a network or just spamming?
11b	Amount of likes	This might indicate how popular the company is, however, it can also be manipulated easily.
12	Other information on Facebook	• Number of content shares. • Reviews about the company.

LinkedIn

Nr.	Question	Results
13b	Linkedin as information source	LinkedIn might reveal some interesting information on network and recommendations of the company / entrepreneur.
14b	Amount of connections	It might give some insight in the amount of business contacts.
15c	Amount of posted messages	The amount of messages says something about how active the company is.
16b	LinkedIn education / experience	This might be interesting to find out the capabilities of persons. For example, higher educated people might be more preferable.
17b	LinkedIn skills	The skills a person have says something about the person's professionalism.
18	Other information on LinkedIn	• Who have seen my profile? • Qualifications which have been verified by others.

Twitter

Nr.	Question	Results
19b	Twitter as information source	It might contain some interesting public announcements.
20b	Amount of followers	This might say something about the influence, reputation and broadcast radius of the company.
21c	Amount of posted / received messages	This might say something about how active the company is.
22b	Message content	It might contain some striking information, for example, complaints and responses to those complaints.
23b	Amount of retweets	This might say something about the popularity of the company and the level of engagement with its customers.
24	Other information on Twitter	How often others Twitter about the company.

25a	Other websites	• Company website. • Google Trends. • Google Reviews: Reviews of customers about the company. • Google Plus. • News websites: The amount of publicity for the company or the classification of the news as positive or negative. • Vacancy websites: The number of vacancies the company has outstanding.

Appendix F

Correlation between Internet features and PD

This appendix shows the correlation between the Internet features of Twitter, Facebook and LinkedIn discussed in Section 5.7 of Chapter 5 and the probability of default of the companies from which these features have been extracted.

F.1 Twitter features

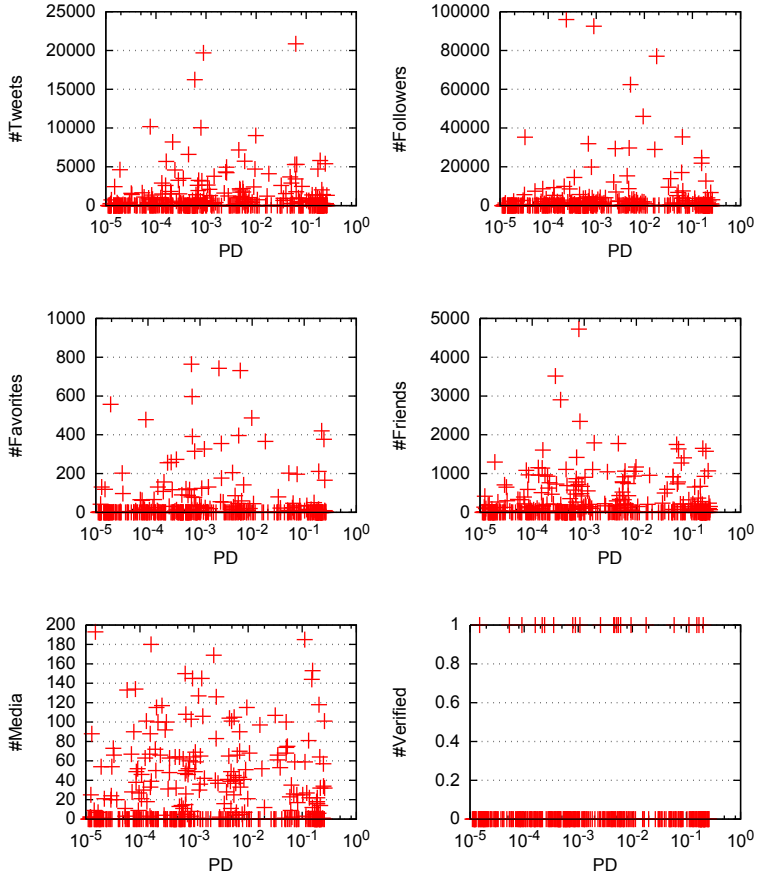

FIGURE F.1: Twitter features plotted against probability of default (PD).

F.2 Facebook features

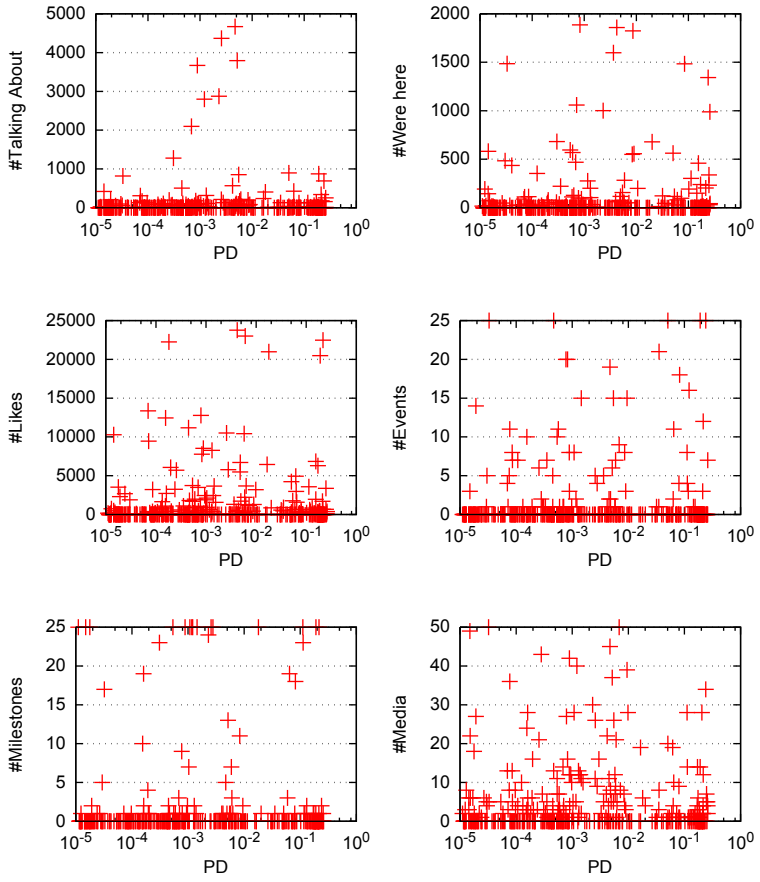

FIGURE F.2: Facebook features plotted against probability of default (PD).

F.3 LinkedIn features

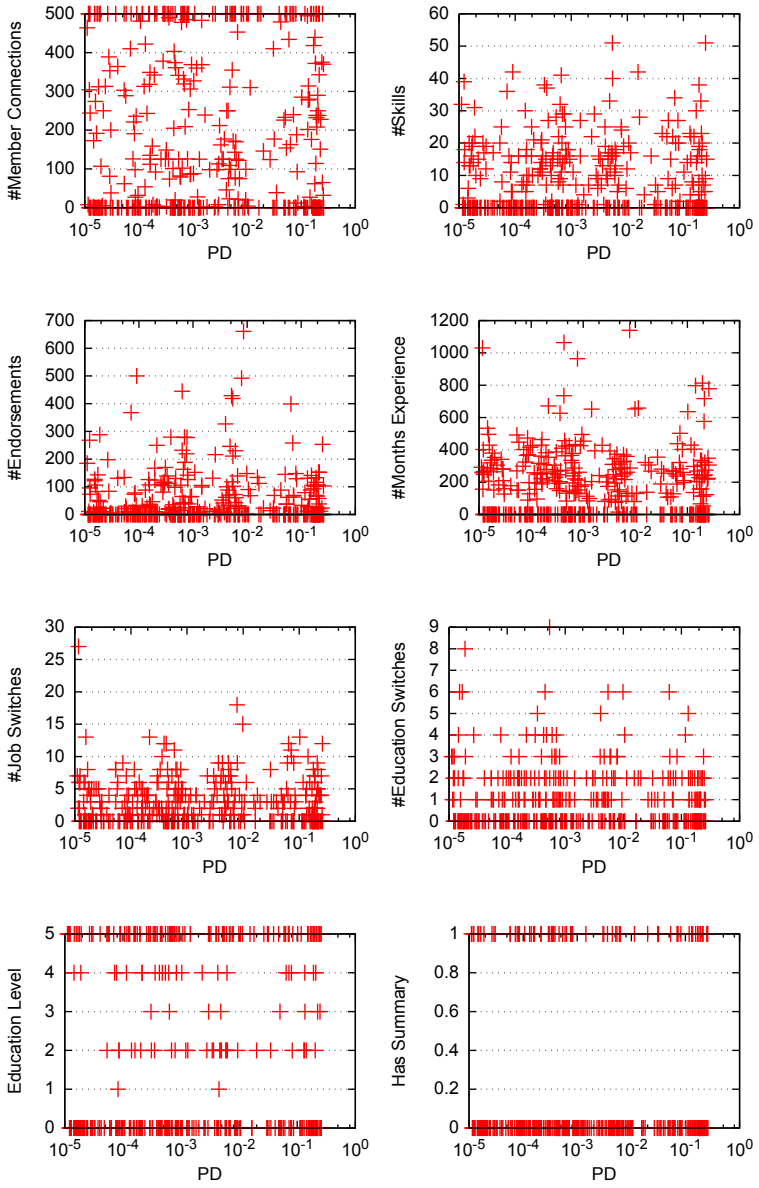

FIGURE F.3: LinkedIn features plotted against probability of default (PD).

Bibliography

[1] K. J. Peeler. The Rise and Fall of J. Pierpont Morgan: The Shift in John Pierpont Morgan's Public Image From the Bailout-out of Moore & Schley Brokerage House in 1907 to the Pujo Hearings in 1913. 2010.

[2] E. M. Rusli. Bad Credit? Start Tweeting. *Wall Street Journal*, April 2013. ISSN 0099-9660. URL `http://on.wsj.com/1zFk8rv`.

[3] Y. Wei, P. Yildirim, C. den Bulte, and C. Dellarocas. Credit Scoring with Social Network Data. *Available at SSRN 2475265*, 2014.

[4] J. Heijnen. *Social Business Intelligence: How and where firms can use social media data for performance measurement, an exploratory study.* Phd thesis, TU Delft, Delft University of Technology, 2012.

[5] L. Dey and S. Haque. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):205–226, 2009. ISSN 1433-2833. doi: 10.1007/s10032-009-0090-z. URL `http://dx.doi.org/10.1007/s10032-009-0090-z`.

[6] B. Hardeman. Lenddo's Social Credit Score: How Who You Know Might Affect Your Next Loan. *The Huffington Post*, 2012. URL `http://www.huffingtonpost.com/bethy-hardeman/lenddos-social-credit-sco_b_1598026.html`.

[7] Toolboek Kredietrisico-analyse: Retail en Private Banking. ABN AMRO, 2014.

[8] Rating. Technical Report September, Bundesverband Deutscher Banken, Berlin, 2010. URL `http://www.dihk.de/ressourcen/downloads/broschuere_rating.pdf`.

[9] A. Henking, C. Bluhm, and L. Fahrmeir. *Kreditrisikomessung: Statistische Grundlagen, Methoden und Modellierung.* Springer-Verlag, 2007.

[10] E. Commission. How to deal with the new rating culture. (July), 2005.

[11] Ura solvency check. Technical report, URA Rating Agency, 2011. URL `http://www.ura.de/images/SC-Englisch-14-09-2011.pdf`.

[12] H. Krehl and A. Fischer. Bilanzratings als Instrument zur Risikofrüherkennung im Prüfungsprozess. In *Perspektiven des Strategischen Controllings*, pages 281–300. Springer, 2010.

[13] H. Been. Finding you on the Internet : Entity resolution on Twitter accounts and real world people, 2013.

[14] M. B. Habib. *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*. PhD thesis, Univ. of Twente, Enschede, 2014.

[15] M. van Keulen and M. B. Habib. Uncertainty Handling in Named Entity Extraction and Disambiguation for Informal Text. In *Uncertainty Reasoning for the Semantic Web III*, volume 8816 of *Lecture Notes in Computer Science*, pages 309–328. Springer Verlag, Berlin, November 2014.

[16] T. Hasanov, M. Ozeki, and N. Oka. Microcredit risk assessment using crowdsourcing and social networks. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014 15th IEEE/ACIS International Conference on*, pages 1–5, 2014. doi: 10.1109/SNPD.2014.6888682.

[17] M. van Keulen. Onzekere databases. *DB/M: database magazine*, 21(4):22–27, 2010. ISSN 0925-6911.

[18] D. Suciu, D. Olteanu, C. Ré, and C. Koch. Probabilistic databases. *Synthesis Lectures on Data Management*, 3(2):1–180, 2011.

[19] N. Dalvi, C. Ré, and D. Suciu. Probabilistic databases: Diamonds in the dirt (extended version), 2008.

[20] M. Magnani and D. Montesi. A Survey on Uncertainty Management in Data Integration. *Journal of Data and Information Quality (JDIQ)*, 2(1):1–33, 2010. ISSN 19361955. doi: 10.1145/1805286.1805291.http.

[21] J. Huang, L. Antova, C. Koch, and D. Olteanu. MayBMS: a probabilistic database management system. *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1071–1074, 2009. doi: 10.1145/1559845.1559984.

[22] J. Widom. Trio: A system for data, uncertainty, and lineage. *Managing and Mining Uncertain Data*, pages 113–148, 2008.

[23] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Re, and D. Suciu. MYSTIQ: a system for finding more answers by using probabilities. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 891–893. ACM, 2005.

[24] S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, and R. Shah. Orion 2.0: native support for uncertain data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1239–1242. ACM, 2008.

[25] M. van Keulen, A. de Keijzer, and W. Alink. A Probabilistic XML Approach to Data Integration. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 459–470, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2285-8. doi: 10.1109/ICDE.2005.11.

[26] P. Stapersma. Efficient Query Evaluation on Probabilistic XML Data. (December), 2012.

[27] B. Qin and Y. Xia. Generating efficient safe query plans for probabilistic databases. *Data & Knowledge Engineering*, 67(3):485–503, 2008.

[28] A. Marouani. Predicting Default Probability Using Delinquency: The Case of French Small Businesses. *Available at SSRN 2395803*, 2014.

[29] J. Melton and S. Buxton. *Querying XML: XQuery, XPath, and SQL/XML in context*. Morgan Kaufmann, 2011.

[30] S. K. Thompson. *Sampling*, page 60. John Wiley & Sons., 2012.

[31] M. Brinkhuis. The Added Value of External Data Sources in the Credit Application Process. 2014.

[32] D. Kedmey. PayPal Co-Founder Takes Aim at Credit Card Industry With New Startup. September 2014. URL `http://time.com/3430817/paypal-levchin-affirm-lending/`.

[33] J. Heijnen, M. de Reuver, H. Bouwman, M. Warnier, and H. Horlings. Social media data relevant for measuring key performance indicators? A content analysis approach. In *Co-created effective, agile, and trusted eServices*, pages 74–84. Springer, 2013.

[34] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and Others. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[35] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2012. ISSN 15324435. doi: 10.1162/jmlr.2003.3.4-5.993.

[36] H. Abdi. Least Squares. *Encyclopedia of Research Design*, pages 1–7, 2010.

[37] R. Gerrits. Social Media Analysis for Credit Rating A Personality Scan of Entrepreneurs by Making Use of Twitter, 2014.

[38] P. Nakra. Corporate reputation management:" crm" with a strategic twist? *Public Relations Quarterly*, 45(2):35–42, 2000.

[39] J. Doorley and H. F. Garcia. *Reputation management: The key to successful public relations and corporate communication*. Routledge, 2011.

[40] Elixir Systems. Online Reputation Management: Protect Your Brand Influence Consumer Perception. Technical report, 2006.

[41] D. Gibson, J. L. Gonzales, and J. Castanon. The importance of reputation and the role of public relations. *Public Relations Quarterly*, 51(3):15, 2006.

[42] R. Dolle. Online reputation management. 2014.

[43] H. Keener. An analysis of online reputation management, 2011.

[44] S. Spencer. DIY reputation management, 2007. URL `http://www.cnet.com/news/diy-reputation-management/`.

[45] S. Kinzie and E. Nakashima. Calling In Pros to Refine Your Google Image, 2007. URL `http://www.washingtonpost.com/wp-dyn/content/article/2007/07/01/AR2007070101355.html`.

[46] T. Krazit. A primer on online reputation management, 2011. URL `http://www.cnet.com/news/a-primer-on-online-reputation-management/`.

[47] European Parliament and Council of the European Union. DIRECTIVE 2005/29/EC, 2005.

[48] N. Kumar and R. N. Reddy. *Automatic detection of fake profiles in online social networks*. PhD thesis, 2012.

[49] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq. Paying for likes?: Understanding facebook like fraud using honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 129–136. ACM, 2014.

[50] X. Lin, M. Xia, and X. Liu. Does" like" really mean like? a study of the facebook fake like phenomenon and an efficient countermeasure. *arXiv preprint arXiv:1503.05414*, 2015.

[51] M. Verma, S. Sofat, et al. Techniques to detect spammers in twitter-a survey. *International Journal of Computer Applications*, 85(10):27–32, 2014.

[52] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM, 2012.

[53] Piers Dillon-Scott. Facebook gets serious about paid brand likes – starts deleting, 2012. URL `http://sociable.co/social-media/facebook-get-serious-about-paid-brand-likes-starts-deleting/`.

[54] Dhiraj Murthy. *Twitter: Social Communication in the Twitter Age*. John Wiley & Sons, 2013.

[55] Ian H. Witten, Eibe Frank, and M. a. Hall. *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0080890369. doi: 10.1002/1521-3773(20010316)40:6⟨9823::AID-ANIE9823⟩3.3.CO;2-C.

[56] B. Qin, Y. Xia, and S. Prabhakar. Rule induction for uncertain data. *Knowledge and information systems*, 29(1):103–130, 2011.

[57] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung. Naive bayes classification of uncertain data. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 944–949. IEEE, 2009.

[58] B. Qin, Y. Xia, and F. Li. Dtu: a decision tree for uncertain data. In *Advances in Knowledge Discovery and Data Mining*, pages 4–15. Springer, 2009.

[59] B. Qin, Y. Xia, R. Sathyesh, S. Prabhakar, and Y. Tu. urule: A rule-based classification system for uncertain data. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1415–1418. IEEE, 2010.

[60] B. Qin, Y. Xia, S. Wang, and X. Du. A novel bayesian classification for uncertain data. *Knowledge-Based Systems*, 24(8):1151–1158, 2011.

[61] J. B. T. Zhang. Support vector classification with input data uncertainty. *Advances in neural information processing systems*, 17:161, 2005.

[62] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *Available at SSRN 2477899*, 2014.

[63] R. Primus. The future of disparate impact. *Michigan Law Review*, pages 1341–1387, 2010.

[64] U.S. Supreme Court. Griggs v. Duke Power Co. 401 U.S. 424, 1971.

[65] U. E. E. O. Commission et al. Title vii of the civil rights act of 1964. *Retrieved February*, 27:2003, 1964.

[66] A. D. United States District Court, N. D. Georgia. EEOC v. Sambo's of Georgia, Inc., 530 F. Supp. 86, 92, 1981.

[67] C. Equal Employment Opportunity Commission, E. E. O. Commission, et al. Uniform guidelines on employee selection procedures. *Federal register*, 43:38295–38312, 1978.

[68] C. O'Neil. Big Data Is The New Phrenology, 2015. URL `BigDataIsTheNewPhrenology`.

[69] J. Podesta. *Big Data: Seizing Opportunities Preserving Values*. 2014.

[70] T. Zarsky. Understanding discrimination in the scored society. *Washington Law Review*, 89(4), 2014.

[71] O. Tene and J. Polonetsky. Judged by the tin man: Individual rights in the age of big data. *J. on Telecomm. & High Tech. L.*, 11:351, 2013.

[72] A. Wynsberghe, H. Been, and M. Keulen. To use or not to use: guidelines for researchers using data from online social networking sites. 2013.

[73] W. B. Persoonsgegevens. Wet van 6 juli 2000, houdende regels inzake de bescherming van persoonsgegevens (wet bescherming persoonsgegevens), 2000.

[74] P. Christen. A comparison of personal name matching: Techniques and practical issues. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 290–294. IEEE, 2006.

[75] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[76] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection*. Morgan and Claypool Publishers, 2010. ISBN 1608452204, 9781608452200.

[77] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293. Association for Computational Linguistics, 2000.

[78] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. pages 73–78, 2003.

[79] J. J. Pollock and A. Zamora. Automatic Spelling Correction in Scientific and Scholarly Text. *Commun. ACM*, 27(4):358–368, April 1984. ISSN 0001-0782. doi: 10.1145/358027.358048. URL `http://doi.acm.org/10.1145/358027.358048`.

[80] W. E. Yancey. Evaluating string comparator performance for record linkage. *Statistical Research Division Research Report, http://www. census. gov/srd/papers/pdf/rrs2005-05. pdf*, 2005.

[81] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[82] M. B. Habib and M. van Keulen. Information Extraction for Social Media. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction (SWAIE 2014), Dublin, Ireland*, volume W14-62, pages 9–16, Dublin, August 2014. Association for Computational Linguistics.

[83] M. B. Habib and M. van Keulen. Unsupervised Improvement of Named Entity Extraction in Short Informal Context Using Disambiguation Clues. In *Workshop on Semantic Web and Information Extraction, SWAIE 2012, Galway, Ireland*, volume 925 of *CEUR Workshop Proceedings*, pages 1–10, Germany, 2012. CEUR-WS.org.

[84] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics, 1997.

[85] M. Asahara and Y. Matsumoto. Japanese Named Entity Extraction with Redundant Morphological Analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 8–15, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073447. URL `http://dx.doi.org/10.3115/1073445.1073447`.

[86] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.

[87] C. C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.

[88] S. M. Weiss and N. Indurkhya. Reduced complexity rule induction. In *IJCAI*, pages 678–684, 1991.

[89] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[90] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[91] S. Sayad. An introduction to data mining. *Data Mining Group, University of Toronto*, 2010.

[92] Y. Yang and G. I. Webb. On why discretization works for naive-bayes classifiers. In *AI 2003: Advances in Artificial Intelligence*, pages 440–452. Springer, 2003.