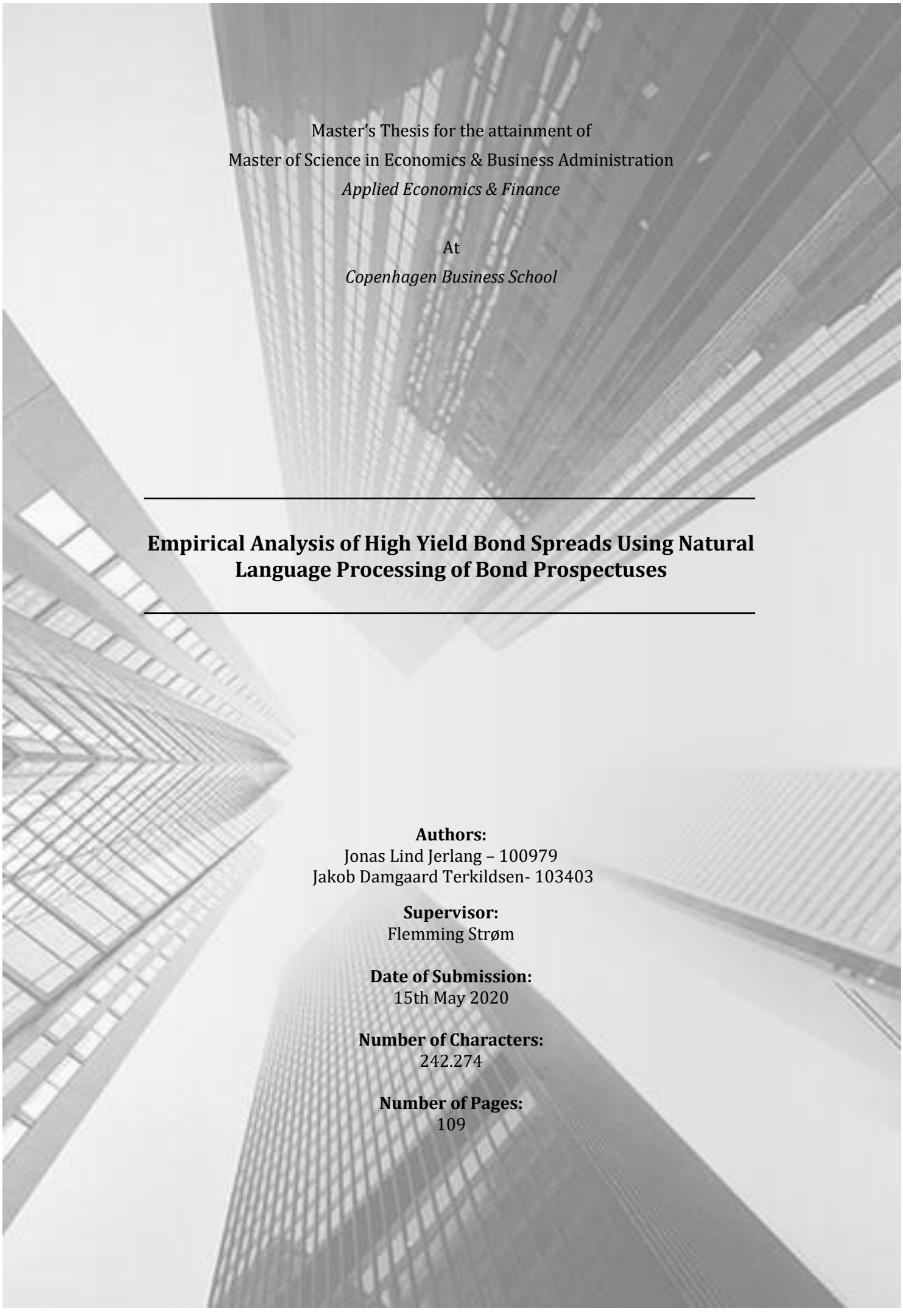


Master's Thesis for the attainment of
Master of Science in Economics & Business Administration
Applied Economics & Finance

At
Copenhagen Business School

Empirical Analysis of High Yield Bond Spreads Using Natural Language Processing of Bond Prospectuses

Authors:
Jonas Lind Jerlang – 100979
Jakob Damgaard Terkildsen- 103403

Supervisor:
Flemming Strøm

Date of Submission:
15th May 2020

Number of Characters:
242.274

Number of Pages:
109

Abstract

In this thesis, we investigate the effect of textual features in bond prospectuses on yield spreads for European high yield bonds. The use of text inputs for security pricing is becoming common practice within equity research, but for the much less developed market of European high yield corporate bonds, inclusion of such unstructured data is unexplored.

Through Natural Language Processing, we extract features from two datasets containing publicly traded European high yield bonds from private and public companies, respectively. These features, combined with traditional accounting variables and bond characteristics, are used to asses yield spreads, first using multiple linear regression, where it is found that textual features of high yield bond prospectuses have statistically significant explanatory power on yield spreads.

We then setup four Machine Learning algorithms to create an analytical framework for yield prediction: A linear Ridge regression, a Random Forest regression, a Support Vector Machine regression, and lastly an ensemble Voting Regressor made by combining the three individual models. When tested on an unseen test dataset, the models have better prediction power of yield spreads when textual variables from bond prospectuses are included, both for private and public firms. While the increase in prediction power is largest for private firms, the models are able to explain a larger proportion of total variance for public firms.

Correctly estimating the weight of textual features when performing security analysis is difficult for analysts, as data is not easily quantifiable or interpretable. An important contribution of this thesis is therefore the development of a framework to quantify these variables and the demonstration that they should be included when modelling yield spreads of European high yield bonds.

Table of Contents

Abbreviation sheet

Abbreviation	Full word
AFME	Association for Financial Markets in Europe
API	Application Program Interface
BofA	Bank of America
BoW	Bag-of-Words
BPS	Basis Points
CAGR	Compounded Annual Growth Rate
CAPEX	Capital Expenditure
CCY	Currency
CDS	Credit default Swap
CUSIP	Committee on Uniform Security Identification Procedures
EBITDA	Earnings Before Interest Tax Depreciation and Amortization
EDGAR	Electronic Data Gathering, Analysis, and Retrieval system
EMH	Efficient Market Hypothesis
EV	Enterprise Value
FOCAS-IE	Feature-Oriented, Context-Aware, Systematic Information Extraction
FV	Face-Value
G-spread	Government Spread
HY	High Yield
ICE	Intercontinental Exchange
ICR	Interest Coverage Ratio
IG	Information Gain
IPO	Initial Public Offering
ISIN	International Securities Identification Number
LBO	Leveraged Buy-Out
LDA	Latent Dirichlet Allocation
LGD	Loss Given Default
LM	Lagrange Multiplier
LRR	Linear Ridge Regression
M&A	Mergers & Acquisitions

MAE	Mean Absolute Error
MD&A	Management Discussion & Answers
nCovid-19	Novel Corona Virus 2019
NI	Net Income
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PD	Probability of Default
PE	Private Equity
PIK	Payment In Kind
PoS	Part of Speech
RBF	Radial Basis Function
Rf	Risk-Free Rate
RFE	Recursive Feature Elimination
RFECV	Recursive Feature Elimination with Cross Validation
RMSE	Root Mean Square Error
ROA	Return On Assets
SEC	US Securities and Exchange Commission
SKLearn	Sci-Kit Learn
SVM	Support Vector Machine
SVR	Support Vector Regressor
T-spread	Treasury Spread
TA	Tangible Assets
TF-IDF	Term-Frequency Inverse-Document-Frequency
VR	Voting Regressor
WC	Working Capital
YTC	Yield to Call
YTM	Yield To Maturity
YTW	Yield To Worst
Z-spread	Zero-volatility Spread

"In the stock market, trading is done electronically by robots. News is instant. Data is abundant. In the world of fixed income, trading is done by humans over the phone, news takes 20 minutes to hit the market, and the data sucks. The debt capital markets behave like it's still the 1980s"

- **Steven Hunter, founder & CEO of 9fin**

1. Introduction

The issuance of bonds with large credit risk has become an increasingly important source of financing for companies operating in Europe, after legislative restrictions were placed on banks following the great financial crisis of 2008. As is the case with stocks, these bonds trade publicly, and analysts attempt to forecast how they will perform in order to make profitable investments. But unlike stocks, the market for high yield corporate bonds is still a relatively young and niche market, only accessible to institutional investors. This implies that both the quality and quantity of data available, as well as the resources to process it, are smaller. Without full information, capital markets are unable to allocate resources optimally, which makes transparency and free information flow in the interest of all actors – both companies, governments, and investors.

Typically, analysts will perform thorough financial analysis of the company in question, with data being obtained from financial records and earnings reports. But, as has been shown in multiple studies performed on equities, text data from news, social media, management descriptions or security prospectuses can contain additional information valuable for pricing financial securities. This data, however, is unstructured. It does not conform to rows and columns, making it more complicated to quantify and model. But with the recent development within Natural Language Processing and machine learning, new ways of extracting information from text related to financial securities are being developed, making it an interesting area of research within the financial literature.

This paper seeks to branch out the analysis of text data for pricing of financial securities, to the universe of high yield bonds. It is written in close collaboration with one of the leading European high yield asset managers, Capital Four Management, as the expansion of this field of research also holds value for industry practitioners. More specifically, the paper takes it onset in the following problem statement:

What determines the yield spread of European high yield bonds, and do the unstructured text data in bond prospectuses contain predictive power of spreads beyond traditional measures such as financial ratios and contractual bond characteristics?

In addition to the overall problem statement, the paper will seek to answer four sub-questions related to the spread of high yield bonds:

- Do accounting ratios and contractual bond characteristics provide information beyond that of credit agency ratings assigned to securities upon the issuance of debt?
- Do textual features of bond prospectuses provide valuable information both at the time of bond issuance and in secondary trading?
- Are there differences in the information content of text data between private and public firms?
- How is information from text in bond prospectuses most effectively captured?

The problem statements will be answered through an empirical study of publicly traded bonds in the European high yield market, with bonds issued between January 1st 2014 and December 31st 2019. The paper employs two separate datasets, one for privately owned companies and one for publicly traded companies. A multiple linear regression is performed, with explanatory variables including bond rating, accounting variables and bond characteristics and textual data from the bond prospectus. Secondly, the paper setup four machine learning algorithms and run the same variables used in the multiple linear regression, to test whether textual features can add the algorithms' explanatory power on an unseen test dataset, which is an analytical setup that could potentially be employed by industry experts such as Capital Four Management, to predict bond performance.

The paper will be structured as follows: A brief introduction to the high yield market will be provided, followed by a theoretical section explaining core concepts behind bonds, with particular focus on the pricing of high yield bonds. Then, an extensive outline of the literature relevant to the study will be presented, with a section covering empirical studies of bond spreads and a section covering the development of Natural Language Processing for the pricing of financial securities. We then consider the theoretical underpinnings of inputs used in the models, before moving to the empirical study, which begins with a section outlining the data collection and data processing performed. The paper then specifies the multiple linear regression model employed, before presenting and discussing the results of the analysis. We then present the theoretical and empirical setup of the four machine learning models employed, after which the results of the models are presented and discussed. Finally, we consider implications for academia and practitioners, including a description of how the analytical setup can be of value to practitioners such as Capital Four Management.

1.1 Acknowledgements and delimitation

The market for European high yield bonds is dominated by institutional investors, with barriers large barriers to enter for retail investors. Minimum trade sizes are in the thousands of euros, and as new issues are often a result of a private equity deal, only certain investors deemed as relevant may have access to key information about the underlying company. As such, access to data on the market is not easily obtainable. We are therefore grateful for the opportunity to write the paper in collaboration with Capital Four Management (Capital Four), one of the leading asset managers within European high yield. Capital Four was founded in 2007 and currently has 11€bn under management, with the biggest mandate coming from Swedish bank Nordea. The client list, however, includes a wide range of investors, from insurance funds to family offices and a long list of Danish pension funds. In 2020, Capital Four was awarded the Lipper Fund Award for best fund over 5 and 10 years in the Nordics within European High Yield Bonds (Capital Four, 2020), a testimony to the longevity of the funds ability to outperform the market.

The data collected for the paper would not have been available to the authors without the resources of Capital Four, and the fact that a large fund manager has shown interest in the development of an analytical framework for text analysis in the pricing of financial securities indicates that the area holds potential, and that it still is not employed by institutional investors in the high yield universe. We are grateful for the support in developing this thesis.

As there are structural difference between the European and the US high yield market, and as Capital Four only operates in the European market, the paper will focus exclusively on bonds issued in European countries. There are several factors affecting the pricing of a bond, and although the paper will provide a detailed outline of these, the study only investigates credit risk, as this is the risk that bond prospectuses are providing information on. Since the contribution of the study concerns the analysis of text in bond prospectuses, the paper will not attempt to invent new measures and ways of analyzing the financial and accounting variables needed to model bond spreads. Instead, it will follow the literature on the subject, and use well-tested and proven methodologies when using inputs from company financial statements.

2. The European High Yield Market

The European high yield bond market is defined as the market for bonds issued by companies in a European country with a credit rating of BBB- or less (S&P) or Baa- or less (Moody's) (Fridson, 2018). The European high yield market is still a relatively young and fast-growing market compared to other European financial markets, such as the market for investment grade bonds or the stock market. The first wave of European high yield took off in the late 1990's, where investors in Europe tried to copy the success of financing the rollout of the telecommunication and media sector in the US through the issuance of high yield bonds (Stone Harbor, 2015). The market was, however, still very infant. In 1998 the market only consisted of €4.9 billion based on 35 issues (Stone Harbor, 2015). This phase of the European high yield market ended poorly shortly after the turn of the century, as the telecommunication became overly invested in and overleveraged with many projects failing to meet the yearly coupon requirements attached to the high yield bond financing structure (Stone Harbor, 2015). The second wave of European high yield bond issues happened in the mid-2000s. This round of growth where also mimicking the U.S. market, and were driven by a heavy wave of leveraged buyouts (also known as Private Equity investments or LBOs), where large amounts of debt was issued to take publicly listed companies private with a highly levered balance sheet. The European high yield market grew in the period of March 2003 to March 2007 from €53bn to €84bn (Stone Harbor, 2015). This phase ended abruptly in 2008 with the crash of the financial markets, which brought an end to the leveraged buyouts of the 2000's.

After a brief slowdown during the peak of the financial crisis in 2008, the growth of the European high yield bond market exploded. Financial troubles of the European banks and new regulatory frameworks which required banks to sharpen their attention to risk exposure and install new capital requirement, constrained the lending appetite of the banks and consequently the ability of companies to finance their investment through the banks. This led many companies to turn to the capital markets for new financing and refinancing, which caused a surge in the number of new European high yield bond issues (Stone Harbor, 2015). Another source of growth in the European high yield market in the post financial crisis period was the resulting number of *Fallen Angels.* Bond issues originally issued as investment grade was downgraded into the high yield space. Since 2009, approximately 27% of the growth in the European high yield market is attributable to fallen angels (Stone Harbor, 2015). By 2015, the European high yield market had more than quadrupled and grown to €387 bn (AFME, 2020). Since then growth has continued, although at a slower pace. As of 31/12/2019, the European High Yield bond market was valued at €508bn outstanding. Figure 1 shows the development of the European High Yield bond market.

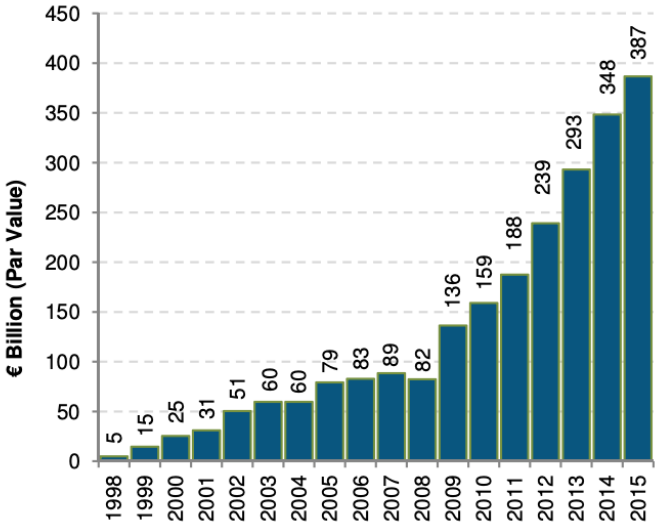


Figure 1: Development of the European high yield market (AFME, 2020)

2.1 Segmentation of the European high yield bond market

Despite the high growth in the European high yield bond market, it remains a relatively low share of the overall European bond market (AFME, 2020). Figure 2 illustrates the value outstanding of European high yield bonds compared to European investment grade bonds. This is partly due to the fact that high yield bond issues are made primarily by privately owned firms, with publicly traded companies preferring investment grade bond or equity issues (S&P, 2020).

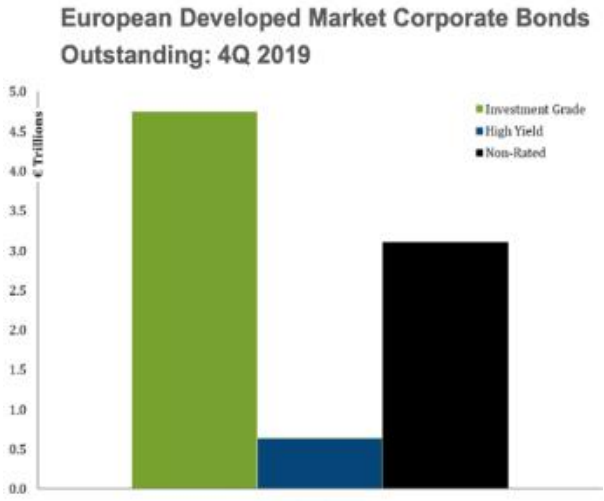


Figure 2: Breakdown of the European corporate bond market (AFME, 2020)

The growth of the market has transformed the European high yield market and greatly diversified the industry, from being almost only telecommunication and media companies to having almost every major industry represented. Figure 3 shows the European high yield bond market broken down by industry. While the communication industry is still heavily represented, the largest industry is by far the financial industry, but also Materials and Consumer Discretionary has grown to take up a substantial part of the European high yield bond market (AFME, 2020).

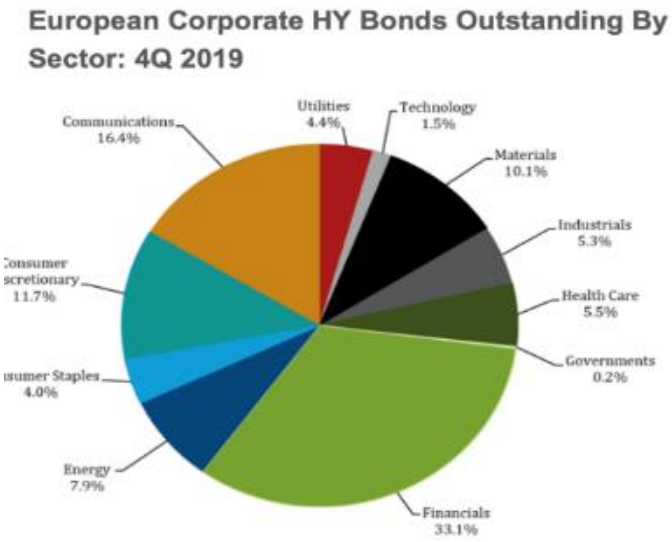


Figure 3: European Corporate HY bonds outstanding by sector Q4 2019 (AFME, 2020)

Figure 4 and 5 shows the breakdown of the European high yield market by credit rating and maturity profile, respectively. The vast majority of the European high yield market are BB ranked, BB+, BB, and BB- making up more than two thirds of the market. The most common maturity

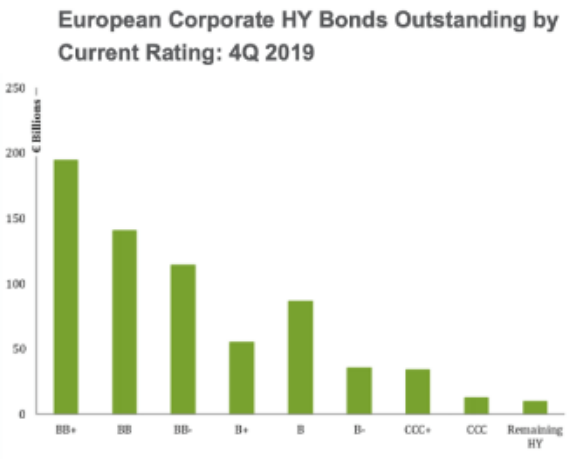


Figure 4: European corporate HY bonds outstanding by rating, Q4 2019 (AFME, 2020)

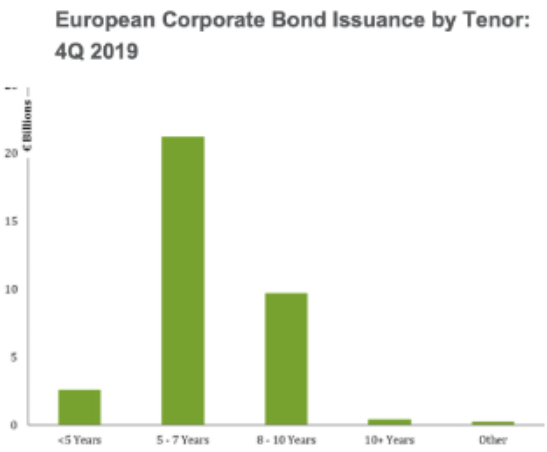


Figure 5: European Corporate bond by maturity profile, Q4 2019 (AFME, 2020)

profile is between 5-7 years, but a substantial part of the market also has a maturity profile of 8-10 years. Only a very small portion of the market have maturity profiles above or below that (AFME, 2020).

Figure 6 shows the use of the proceeds raised in European high yield bond issuances. Over the last 10 years, the main use of proceeds is General Corporate Purposes, such as investments in PP&E or general CAPEX and development of the issuing company. A substantial part of the proceeds is also applied towards refinancing of debt. This indicates that a substantial part of the European high yield market is keeping high levels of debt in the capital structure on a more permanent and strategic basis. Another significant use of proceeds is Acquisitions. This is a more volatile use and follows the overall M&A activity level in Europe (AFME, 2020). Leveraged buyouts have gained some popularity again (Gottfried, 2018), but remains a quite small part of the total use of proceeds.

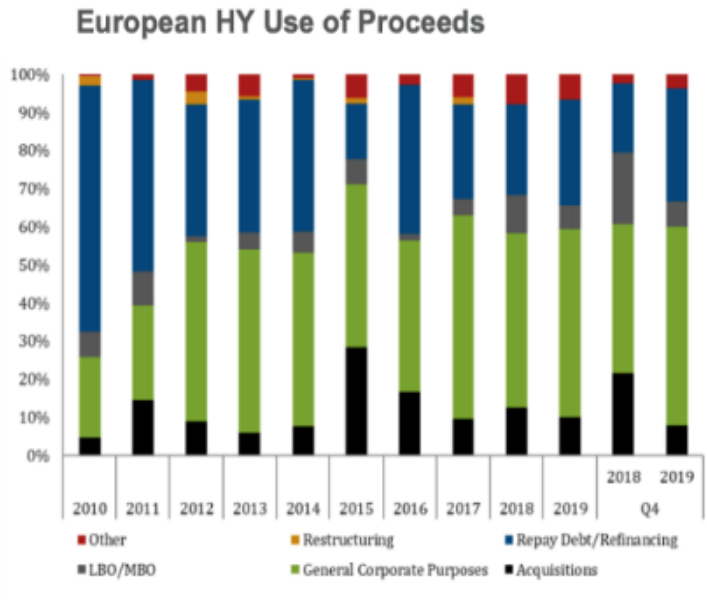


Figure 6: European HY bonds Use of Proceeds (AFME, 2020)

2.2 Risk and return profile of European high yield

The risk profile of European high yield bond differs greatly from the European investment grade bond market. Whereas Investment grades bonds' risk profile mainly comes from interest rate risk and liquidity risk, and possess very little credit risk from potential defaults, with an average default rate of 0,03% (S&P, 2017), high yield bonds hold substantial credit risk with an average

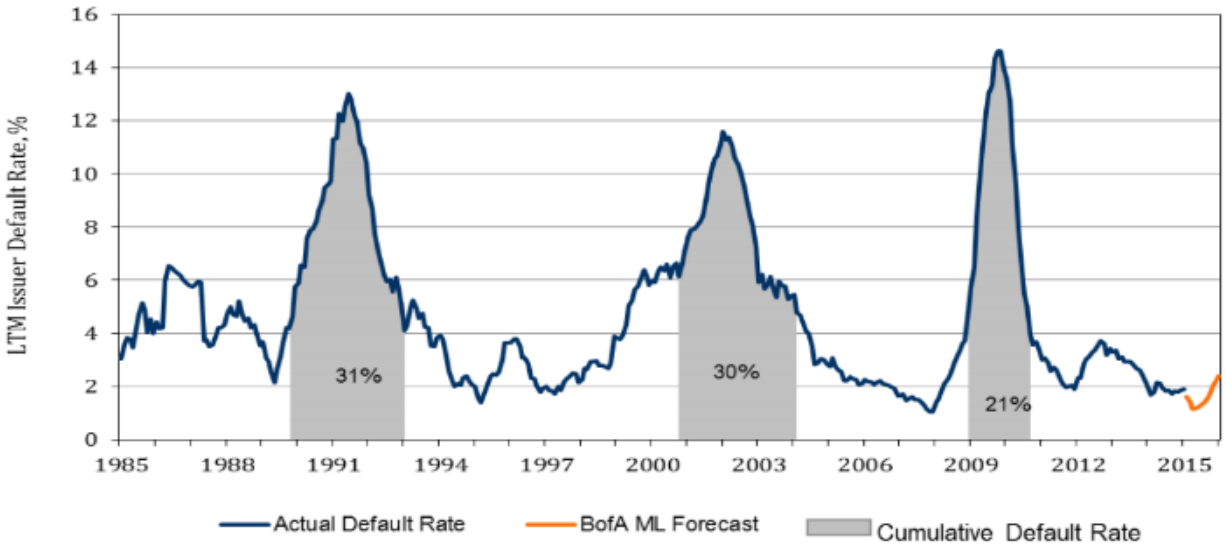


Figure 7: Historic annual and cummulative default rates of European HY bonds (BofA Merril Lynch, 2016)

default risk of 3,1% (S&P, 2019). Figure 6 shows the historic annual and cumulative defaults on high yield bonds. The figure shows how the default rate of European high yield bonds is highly volatile and how high yield bonds are sensitive to the economic cycle and large part of the cumulative defaults accumulate during economic crises, such as the one in the early 90's the dot-com crisis of 2001 and the financial crisis of 2009. The default rates peaked in 2009 following the financial crisis with a default rate of more than 14%.

Figure 8 shows the spreads of European high yield and investment grade bonds and the corresponding default risk across European corporate bonds. It shows that European high yield bond spreads are highly correlated with corporate defaults, with default rates trailing the changes in the high yield bond spread with around a year. Actual defaults often take time to realize and the lag on default rates shows how investors are able to expect them a little ahead and price them into the spread of bonds (BofA Merril Lynch, 2016). Investment grade bonds on the other hand, is only weakly correlated with the default rates, and only rises with default rates in extreme cases such as the financial crisis of 2009, where default rates crept all the way up to the investment grade tranches of corporate bonds (BofA Merril Lynch, 2015).

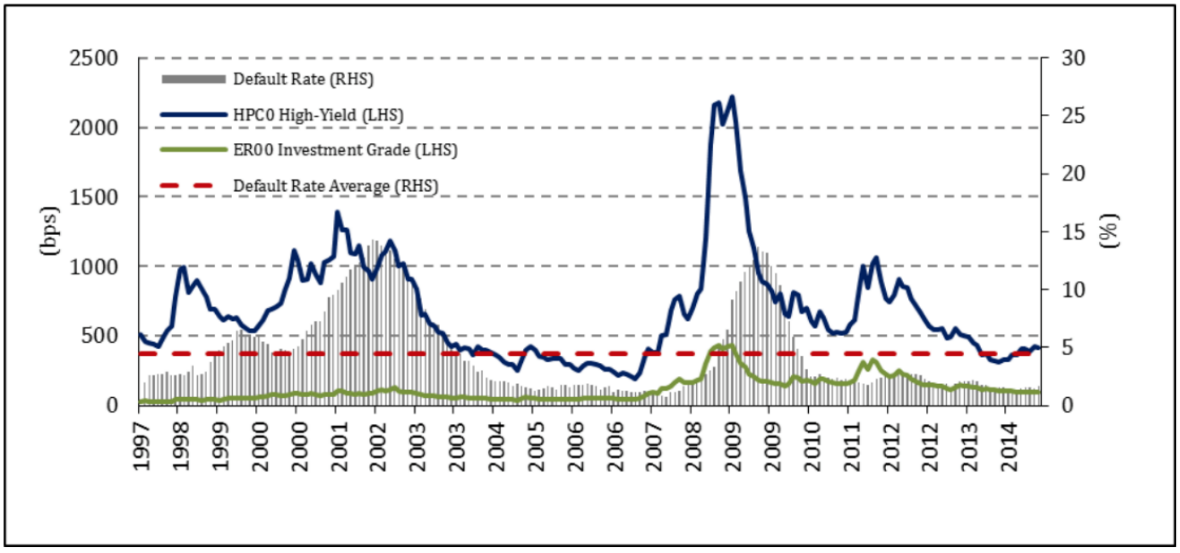


Figure 8: European HY and IG bond spreads and corporate default rates (BofA Merril Lynch, 2015)

Lastly, while the European high yield bond market is an often-overlooked asset class, that does not attract nearly as much attention as the stock market or money market, it has performed extremely well historically. As figure 9 shows, the European High Yield bond market has outperformed both the European equity market and the investment grade bond market over the last 19 years, with a compounded annual growth rate (CAGR) of ~7,5%. This outperformance has, for a large part, materialized in the last 10 years, in the period after the 2008 financial crisis (Alfawise, 2020; Bloomberg, 2020; ICE Dataservices, 2020). This high performance makes European high yield bonds as an asset class interesting for research.

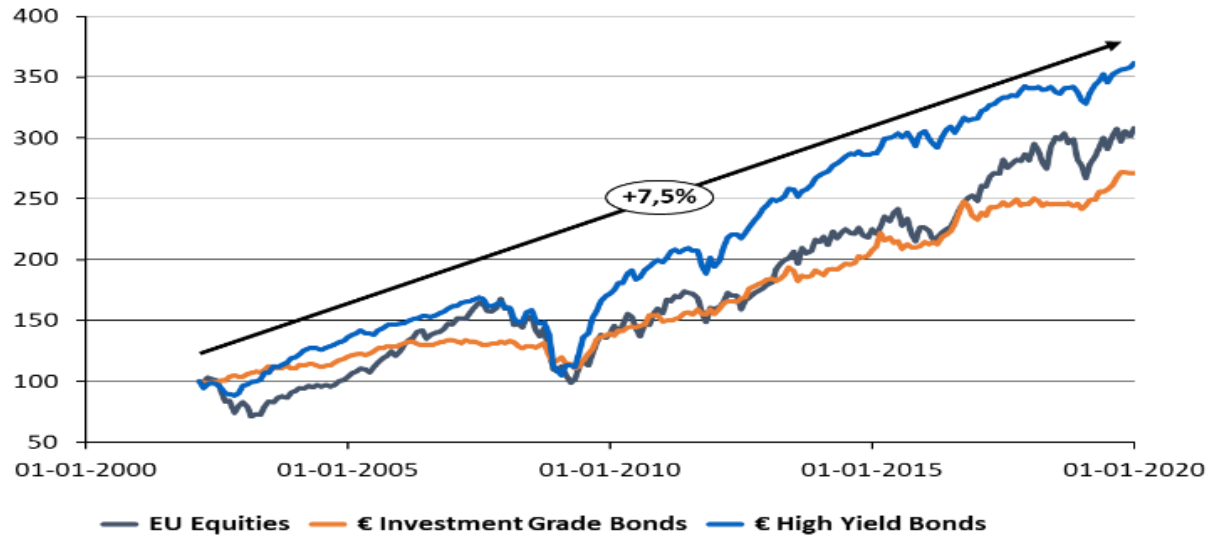


Figure 9: The relative performance of Euro denominated HY bonds 2002 - 2019 (Alfawise, 2020)

3. Theory

With a brief introduction to the market in question provided, the following section outlines the theoretical perspectives relevant for understanding how high yield bonds function, and introduces the financial theory used in the pricing of these securities.

3.1 Bond Basics:

A bond is a type of debt instrument where the issuer of the bond receives a direct inflow of cash upon issuance and is followingly required by contract to repay the lender/investor over the contractual period of the bond plus additional interest payments. A bond differs from other types of debt instruments, such as bank loans, in that all bond share some standardized contractual features which makes them much more easily tradeable on financial markets (Fabozzi, 2013). Globally, there are hundreds of thousands of bond issues (Fabozzi, 2013). The majority of bonds are issued with a nine-character tag for identification called CUSIP (Committee on Uniform Security Identification Procedures). This CUSIP tag allows for precise identification of the specific bonds when traded on the financial markets, especially when a single issuer can have many different bonds outstanding. An alternative and globally used form of identification to the CUSIP tag is the ISIN number (International Securities Identification Number), which similarly assigns a code that uniquely identifies a specific security issue. Throughout this paper, the ISIN numbers will be used to keep track of the specific bonds and the underlying data that refers to the specific bond.

3.1.1 Types of bonds

Many different types of bonds are issued worldwide. One way to distinguish between different types of bonds is by the type of issuer. The issuers of bonds are plentiful and diverse, but the three

main types of issuers are: Governments, municipalities, and Corporations. Government bonds include treasury bills, notes and bonds (the difference being the time to maturity of the bonds), and the rates on these are fundamental for the levels of interest rates around the world (Fabozzi, 2013). Municipality bonds are issued by local government and authorities to raise funds. These bonds are characterized by, in the US, the returns for the largest share of the market being tax exempt. Lastly are the corporate bonds, issued by companies to raise capital for their operations or for new investments. In some countries, there is also large asset-backed bond markets. As an example, both the US and Denmark have a large bond market for mortgage backed bonds (Brealy et al, 2015). This paper will solely focus on corporate bonds.

Another way to distinguish between different types of bonds are through the cash flow structure of the bond. The amount outstanding can either be paid back at the time of maturity, or at any contractually given rate throughout the course of the bond. The most common corporate bond is a bullet bond, where the principal amount is paid back at maturity and interest are paid semi-annually. Interest payments can either be a fixed rate or a floating rate. For fixed rated bonds' interest is set as a fixed percentage of the principal amount. For a floating rate bond, the percentage of the principal required in interest payments are linked to some reference rate, e.g. LIBOR, and typically takes the form of a spread on top of the reference rate. A somewhat common alternative to the typical bullet bond amongst high yield bonds are Pay-In-Kind bonds (PIK). At the time of interest payment, a Pay-In-kind bond gives the issuer a choice of two different types of interest payments. Either the issuer can pay an interest payment in cash or they can pay a slightly higher interest rate but in the form of more debt rather than in cash. This reduces the interest burden and the cash flow constraints on the company, which can be attractive for highly levered companies. PIK bonds are often issued for projects where cash flows during the initial lifespan of the project are very uncertain, or more commonly by companies that are highly levered and with high amounts of debt outstanding relative to their earnings. The downside for PIK bonds as a type of financing is that the cost of capital is higher. The cost of capital is higher due to two theoretical causes. First the actual cash flow of the interest payment falls further into the future, requiring the PIK rate to compensate the investor for the time value of money for that cashflow deferral. Furthermore, giving companies and managers the option not to pay for debt through cash provide the manager with more slack and increases the probability of corporate governance issues and misalignment of interests between the bond investors and the owners of the company (Hart, 2001). This increased corporate governance risk would require a higher interest rate for both the cash rate and the PIK rate.

3.1.2 Corporate bond pricing

The following section will lay out the theory pricing bonds that is necessary to understand when trying to predict price movements or performance of bonds in the corporate bond market.

A basic common bullet bond can be priced using the present value formula:

$$P = \sum_{i=1}^{n} \frac{C}{(1+r)^t} + \frac{FV}{(1+r)^n}$$

Where P is the price of the bond, C is the coupon payment, FV is the face value of the bond, n is the number of periods until maturity and r is the discount rate.

The formula states that the price of a bond is equal the sum of the present value of all cash flows in the bond contract. In reality the prices are easily observed in the market and the unknown factor is the effective interest rate that makes the present value of the cash flows equal to the price of the bond, known as YTM (Yield To Maturity). In the equation above, YTM is the r that makes the present value of the cash flows equal the price P.

3.1.3 Bond returns

For a risk-free bond, the expected return will be exactly equal to the YTM, and for a perfectly liquid risk-free bond, the only factor that can change realized returns to something different than expected returns are movements in the risk-free rate. In this paper the unit of analysis is, however, high yield bonds, which are not risk free since they hold substantial default risk. For corporate bonds, and especially high yield bonds, there is a risk that the issuer will not be able to meet its obligations and will have to default on the payment obligations of the bond. Thus, the expected return on a corporate bond can be estimated as (Fridson, 2018):

$$Expected\ Return = \ Initial\ Yield - Default\ Loss\ Rate$$

Because corporate bonds have a risk of defaulting, the yield paid on corporate bonds has to be higher than on risk-free instruments to both compensate for the direct negative effect of the default loss rate on expected return and for the increased variance / risk it provides (Berk & DeMarzo, 2017). Therefore, corporate bonds are said to trade at a spread above risk-free instruments. This spread is called the *credit risk spread*, *the yield spread* or simply *the spread*. In this assignment the terms for the credit risk spread will be used interchangeably. The credit risk spread is defined as the YTM implied from the market less the risk-free rate:

$$Credit\ Spread = YTM - Rf$$

Consequently, the expected return formula for corporate bonds can be expanded as the following (Fridson, 2018):

$$Expected\ Return = (Rf + Credit\ Spread) - (Default\ rate + Recovery\ rate)$$

Many researchers, such as Huang & Huang (2003), Huang & Huang (2012), Longstaff et al (2005), Ericsson & Renault (2006), and Dick-Nielsen et al (2012) have found the credit spread to mainly come from two source: *Credit Risk* and *Liquidity Risk.* Both measures are explained in greater detail in the Credit Spread section.

3.2 Structural and contractual considerations of corporate bond investing

The following section will lay out the most important and common structural and legal features of bonds and the bond market that can affect the price and performance of corporate bonds.

3.2.1 Seniority and Security

One of the most important structural features of a bond to consider when evaluating the risk of the asset is the seniority and the security of the bond. Seniority refers to the priority right the security holds on cash flows in the case of any liquidity issues. In general, bonds are referred to as *Senior* or *Subordinated*[1]. In the case of default or a company having trouble to meet all its debt obligations, senior ranked bonds are to be paid in full, both interest and principal, before anything are to be paid to the holders of subordinated instruments (Berk & DeMarzo, 2017). That means that, ceteris paribus, the risk of a senior bond is lower than that of a subordinated bond, as more value would have to wiped off the company, before the senior bond holders are not paid in full, and thus the recovery rate will be higher. This is also seen empirically, as historically senior secured bonds have had higher recovery rates than subordinated debt (Fabozzi, 2013). To compensate for the increased risk of subordinated debt, investors will demand a higher risk premium. Secured debt is debt backed by or secured against some form of collateral beyond the issuers general credit standing (Fabozzi, 2013). In the case of default, the value of that collateral is claimed in full to cover the credit obligations towards the secured note holder, and any residual value from the collateralized assets will not fall to other creditors before the secured note holder is paid in full. In the case that the collateral will not cover the full amount outstanding on the bond, the rest of the claims attributed to the bond will be claimed according to its seniority rank. Similar to seniority, investors will demand a higher risk premium for unsecured debt compared to secured debt.

In general, the priority of the debt structure is (Fabozzi, 2013):

[1] Subordinated bonds are also referred to as Junior

- Senior Secured notes
- Senior unsecured notes
- Senior subordinated notes
- Subordinated notes

In the bond market and bond literature other linguistics may be used to describe seniority and security of bonds. *1st Lien* are used to describe the debt holds priority in the debt structure over anything else. In some issues, 1st Lien is used instead of senior debt and in some cases, there will be a class of 1st Lien debt that holds priority over the senior ranked debt. Similarly, *2nd Lien* are used interchangeably with subordinated debt. When two debt instruments claims have the same ranking, they are said to be *pari passu*. In the case of a default where the full claims of two bonds ranking pari passu cannot be met, the creditors will split the payments in proportion to their claims.

In the dataset used in this study, the effect of seniority on credit risk spreads is controlled for, grouping bonds in two categories: Senior, comprising senior secured notes, senior unsecured notes and senior subordinated notes, and Junior, comprising subordinated notes.

3.2.2 Covenants & Provisions

Bond holders are receiving fixed payments in accordance with the contractual claims of the bond, while equity holders have a claim on the residual values that are left, only after the bond holders' claims are paid in full. In many cases, this difference in payments claim can lead to misalignment of the interests of equity holders and bond holders (Laeven & Levine, 2008). For example, in the case of a near term default, equity holders will prefer actions with high risk and high rewards, as it can reward them with high returns in positive outcome and mainly wipe off value of the debt holders in negative outcomes. As equity holders are in control of the company, certain contractual restrictions on management and legally enforceable rules are written into the bond contract. These contractual indenture provisions are commonly referred to as *covenants* and are safeguards for the bondholders against misalignment of interest. In general, there are two types of covenants: *Positive* or *Affirmative covenants* require the issuer to take certain actions, e.g. sell of part of the business to deleverage, make certain investment or change management etc. *Negative* or *restrictive covenants* prohibit the company from doing certain things, e.g. undertake more M&A, make certain investments, take on more debt etc.

The following is a brief introduction to some most common covenants in the bond market. One of the most common covenants in bond contracts is limitations on indebtedness. This indenture specifies either some limits to the absolute level of debt outstanding, or some ratios for which the company's debt level must comply to. Under limitations of indebtedness, if a corporation wants to take on more debt, it must pass a debt incurrence test (Fabozzi, 2013). Two of the most commonly

used ratios for such tests are leverage (Net Debt / EBITDA) or Interest Coverage Ratio (EBITDA / Interest Expense). A similar common covenant is limitations on liens, which protect the bondholders' position in the debt structure. This prohibits the borrower for pledging its assets as collateral in any way that alters the security of the bondholders' bond or issuing new debt that are ranked senior to the bondholders' debt. Another type of covenant, limitations to cash outflows or payments, are a covenant designed to protect the coverage of bondholders. Such limitations restrict the company from paying out any excess cash as dividends or spending it on Investments or in M&A activity. Under this indenture the company is typically restricted from spending cash on such activities as long as it does not live up to certain tests, such as leverage or interest coverage ratio tests, thereby preventing equity holder from spending cash if this would put the claims of the bondholders at any substantial risk.

Because of the protections that covenants can provide, they are important aspects of the risk profile of bonds. Because they can affect risk, they are likewise important determinants of the spreads and prices that bonds will trade at (Fabozzi, 2013).

3.2.3 Other contractual provisions

The following are other contractual provisions that are common amongst corporate bonds:

3.2.3.1 Callable provision:

A call provision in a bond contract, provides the issuer with the right to pay back the bond at a contractual given price before the maturity date. This price is often equal to the face value of the bond plus a predetermined call premium that will compensate the investor for lost interest. Often a bond will have a certain grace period after issuance where the bond is not allowed to be called and after the grace period the bond can be called on certain dates or intervals following a *call schedule*. Most common grace periods are 2, 3 or 5 years after issuance. A callable provision for the bond provides three downsides for an investor (Fabozzi, 2013): Firstly, a callable provision introduces uncertainty of the timing of the cash flow structure. Secondly, as borrowers are assumed to be rational, they would often refinance and use the call option if the interest rate environment has declined, introducing reinvestment risk to the investor. Thirdly, the potential upside for price appreciation is limited, as investors would be hesitant towards investing in the bond at prices above the call price as it would be likely be called in such a scenario. As a consequence, investors require a discount in the price of a callable bond (a spread premium) compared to an identical non-callable bond. The price discount required will be exactly equal to the value of the call option (which can be calculated using any option valuation technique such as the Black-Scholes model), and can be calculated using the following formula (Bodie et al, 2017)

$$Price_{callable} = Price_{Vanilla\ bond} - Price_{Call\ option}$$

The introduction of a call option also complicates the yield calculations on the bond. For a callable bond, Yield to Maturity will no longer be the only important yield measure. The bond will now

also have a *Yield to Call (YTC)*, which applies if the company chose to call it on a certain date. Yield to call is not calculated any different than Yield to Maturity (YTM) and is simply the interest rate that will make all the future cash equal to the price when discounted. But because the cash flow structure is different if the bond is called, the yield to call will potentially be different than the yield to maturity, depending on the price of the call and the timing of the cash flows. With callable bonds there will be a potential cash flow structure and Yield to Call for each call date in the call schedule. This problem can be circumvented by calculating what is known as *Yield to Worst (YTW)*. The yield to worst is simply the lowest yield of the YTM and all the different YTCs. If the bond is not callable the YTW will be identical to the YTM. Under the assumption that the borrower is a rational actor, she will exercise the call option if the price rises above the call option, and the expected yield on the bond will therefore be equal to the YTW. Most high yield bonds (and corporate bonds in general) have a callable provision. In the dataset used in this study this is also the case. Therefore, to use measures that are directly comparable across bonds, yield to worst will be the yield measure applied throughout the paper. Additionally, differences in spreads between bonds with and without call provisions will be controlled when setting up the models.

3.2.3.2 Puttable provision:

similar to a callable provision a bond can hold a puttable provision. While a callable provision gives the issuer the opportunity to redeem the bond at a certain date and predetermined price, a puttable provision gives the investor the opportunity to force the issuer to redeem the bond at a predetermined price and date. This provides the investor with increased control over the cash flow structure while increasing the uncertainty for the issuer. As a consequence, the issuer will demand a discount in the yield she has to pay on the bond. The price can be calculated as (Berk & DeMarzo, 2017):

$$Price_{puttable} = Price_{Vanilla\ bond} + Price_{put\ option}$$

Similarly, the YTW can work for bonds with put provisions as well, by including the yields in the put scenarios in the search for the worst yield. No bonds in the dataset used in this study contain a puttable provision.

3.2.3.3 Convertibility provision:

a convertibility provision is a conversion right that provides the bondholder with the right to convert the face value of the bond into a specified number of common or preferred stock. The underlying price of the stocks in the option will be equal to the face value divided by the number of stocks, also referred to as the conversion price. It is common for issuers to issue bonds with convertibility provisions that are way out of the money, meaning that price of the stock will be lower than the conversion price, making the conversion attractive only in the case of severe appreciation in the stock price. A convertibility provision in a bond thereby allow the investor capture some of the upside potential in the company. Consequently, a convertible bond can be thought of and priced as a regular bond plus an out of the money call option (Berk & DeMarzo,

2017). As a majority of high yield issues in Europe are done by privately held companies, convertibility provisions are not very common in this market.

3.2.3.4 Special structures for high yield bonds:

In the earlier days of the corporate high yield market, high yield debt was either companies with very secure cashflows that chose to lever up highly, or fallen angels, companies that originally issued the debt as investment grade but have been downgraded into the high yield territory. These early bonds had conventional structures with fixed terms and coupon rates (Fabozzi, 2013). Today, however, many bonds are issued by companies that have been taken over by private equity (PE) firms through leveraged buy outs (LBOs) with much more complex structures. In LBOs or PE deals, the level of debt is often set at a very high level to maximize the potential return on equity. Under normal structures, such high levels of indebtedness would place severe cash flow constraints on the company, in order to pay its interest. To reduce the cash flow constraints of the interest burden many of these bonds have been issued with different kind of *deferred coupon structures* (Fabozzi, 2013). Deferred coupon structures let the issuer avoid paying the interest through cash for a certain period. The three most common of such structures are: deferred interest bonds, step-up bonds and payment-in-kind bonds. Deferred interest bonds sell at a very steep discount but pays no interest in the first few years. Step-up bonds do pay interest, but often a low initial interest that then increases over time at certain step-up dates. PIK bonds, as described earlier are the most common, which allows the issuer to choose between a cash interest and a slightly higher non-cash interest (Fabozzi, 2013).

3.3 Risk spreads of high yield bonds

Having laid out the technicalities of corporate bonds, including basic features and characteristics, the basic pricing mechanisms, as well as structural and contractual features that corporate bonds exhibit, the following section will investigate the unique risks associated with high yield bonds and the yield to compensate for this.

3.3.1 Credit spread

As mentioned in the pricing section, corporate bonds, particularly high yield bonds, hold some inherent risk compared to risk-free bonds. To compensate for this risk, high yield bonds trade at a spread, often referred to as the credit spread, above the yield given by risk-free assets. $Credit\ Spread = YTM - Rf$. Many scholars have investigated this spread[2], and while they have different results of the mix between the two, they are in accordance that most of this spread can be explained by two different sources: *Credit risk* and *liquidity risk*. The following section will lay out the basic theory behind these concepts, followed by an overview of the literature related to each of the two.

[2] See section on literature review

With the yield spread defined as being composed of credit risk and liquidity risk, the yield of high yield bonds can be decomposed to the following formula:

$$Yield = R_f + Credit\ risk\ premium + liquidity\ risk\ premium$$

This shows that the yield, and thereby the prices, of high yield bonds are driven by changes in the risk-free interest rate environment, changes in the underlying credit risk of the asset, and changes in the liquidity of the high yield bond market. The effect of changes in the risk-free interest environment should not affect high yield bonds any different than risk-free bonds of similar cash flow and maturity structure. This risk has been investigated in great mathematical detail over many years and can easily be hedged by investors in the construction of their portfolio (Bodie et al, 2018). Furthermore, it should not affect the credit spread of the bonds, which is the unit of analysis in this paper. For these reasons, the paper will not dive deeper into the effect of interest rate risk on high yield spreads. As for liquidity risk, in general, the corporate high yield market is far from perfectly liquid. Bond issues are made in very large amounts[3] directly to institutional investors, who need contacts with large European banks and a solid track record to be considered for a deal. The market is essentially not accessible to retail investors (Bodie et al, 2018). Consequently, investors require a premium to compensate for this illiquidity. While the liquidity share of the overall credit premium has been found to be very significant for corporate bonds in general, but due to the high risk profile of high yield bonds, it has been found to only be a minor share of the credit spread for high yield bonds (Huang & Huang, 2003; Dick-Nielsen, 2012; Lin et al, 2011). For this reason, the paper will be focused on the credit risk inherent in high yield bonds, which the following section will describe in greater detail.

3.3.2 Credit Risk

Credit risk is the risk of an issuer defaulting on some of the obligations of the bond, meaning that the investor will not receive payment in full on all of the claims laid out in the bond. More specifically, Fabozzi (2013) defines credit risk as *"the risk that the issuer of a bond will fail to satisfy the terms of the obligation with respect to the timely payment of interest and repayment of the amount borrowed"*. A common measure used to understand the risk of default on a bond is expected loss, which allows for analysis in greater detail by breaking down the credit risk further.

3.3.2.1 Expected Loss

When a company's earnings and cash flow deteriorate to a level where it is no longer able to serve the obligations of its debt outstanding, it will have to default on its obligations. When that happens, the debtholders will legally take over the rights of the company and proceed to sell off the company, or part of its assets to service the debt (unless a restructuring is negotiated). The expected loss on a corporate bond thus consist of two subcomponents: the probability that the company will have to default on its obligations (the default rate), and the percentage rate of the

[3] The average issue size in the dataset used for the study is around 500€ million

value of the bond that can be recovered when the bondholders legally take over the company (the recovery rate), and can be summed up as:

$$Expected\ Loss = probability\ of\ default * (100\% - recovery\ rate)$$

The recovery rate will mainly depend on the *tangibility of assets*, that is to what degree company assets are tangible, and therefore easy to sell to a third party, but also on the amount of potential buyers in the market that would be interested in bidding for the company part of its assets. To understand the probability of default of a bond, one would need to understand the characteristics of the underlying company. Fabozzi (2013), breaks down the analysis of the probability of default into three underlying types of risk: Business Risk, Corporate Governance Risk, and Financial Risk. Business risk concerns the operating cash flows of the company. This means the trends, opportunities and risk of the company and the industry it operates in. The rating agencies describe the areas of analysis under this parameter as: country risk, industry trend and characteristic, competitive position, product portfolio, strategic and operational management competencies and peer group comparisons (Fabozzi, 2013). Corporate governance risk refers to the risk of misalignment of interest and principal-agency problems that may arise between the bond holders and either the management of the company or the owners / equity holders (Laeven & Levin, 2008). As mentioned earlier, some of these risks can be mitigated through strong protective covenants in the bond contract. Financial risk assessment is the assessment of the direct risk that the company's financial position will be too weak to meet the requirements of the bond obligations. It involves ratio analysis on the financial metrics and common analysis is ratios such as Interest Coverage Ratio, Leverage, Cash Flow Analysis, Margin analysis and Net Asset composition (Fabozzi, 2013).

3.3.3 Credit ratings

For most large issues, the credit risk of bond is assessed by three major credit rating agencies: Standard & Poor Corporation (S&P), Moody's Investor Service (Moody's), and Fitch Investor service (Fitch). These rating agencies provides a relative assessment of the credit risk of the issuer and its ability to pay back the bond timely over the maturity of the bond. Based on this assessment they assign a letter grade to each bond they cover. The below table summarizes the rating scale used by S&P and Moody's. S&P further modify the letter ratings by + and – and Moody's with the numbers 1,2,3 to reach higher granularity in the ratings:

	Moody's	*S&P*	*Description*
Investment grade	Aaa	AAA	Judged to be off the highest quality, subject to the lowest level of credit risk.
	Aa	AA	Judged to be off high quality and are subject to very low credit risk.
	A	A	Judged to be upper-medium grade and are subject to low credit risk.
	Baa	BBB	Judged to be medium-grade and subject to moderate credit risk and as such may possess certain speculative characteristics.
High Yield	Ba	BB	Judged to be speculative and are subject to substantial credit risk.
	B	B	Considered speculative and are subject to high credit risk.
	Caa	CCC	Judged to be speculative of poor standing and are subject to very high credit risk.
	Ca	CC	Highly Speculative and are likely in, or very near, default, with some prospect of recovery of principal and interest.
		C	
Default	C	D	The lowest rated and are typically in default, with little prospect for recovery of principal or interest.

Figure 10: Moody's and S&P's Credit ratings with descriptions (Fabozzi, 2013)

However, several researchers have found that credit spreads can vary a great deal within the same ratings (Fridson et al, 2016; John et al, 2010). For instance, Fridson et al (2016) found that the spread of subordinated debt and senior debt of the same ranking trade at very different spreads. John et al (2010) also found great variance in the spread of different similar ranked credits. Furthermore, as ratings are publicly available to all, they are not a great source of information for fond managers to gain a competitive advantage to beat the market. For researchers, the lack of transparency in the exact calculations that make up the final credit make them unideal as a model to explain the credit risk spread, and thus several researchers have tried to create models that can explain the credit risk spread (Fisher, 1959; Altman, 1968; Merton 1974; Fridson & Garman, 1998; Longstaff et al, 2005; Fridson et al, 2016). Many of these have used structural models and ratio analysis, with a strong focus on the financial risk from the above framework. While the exact model and ratio used varies from researcher to researcher variables covering the following five areas should be applied: liquidity, profitability, leverage, solvency, and activity (Altman, 2000).

Traditionally, the data used to describe these five areas will be of purely quantitative nature, taken from the financial reporting of the company in question. However, important risk such as corporate governance risk can be hidden in the covenant section of a bond prospectus, and otherwise unknown underlying risk, opportunities or management thoughts can be found in the

risk section or management discussion and answers (MD&A) section of a bond prospectuses. As such, the addition of textual features to the traditional financial analysis could provide additional information on one or more of the five areas, and thereby allow investors to make better assessments about the default risk and recovery rate of the underlying company, thus making more accurate estimations of the credit risk.

With the features commonly associated with high yield bonds introduced, and the theoretical underpinnings behind bond pricing and spread calculation introduced, we turn to an examination of previous work related to the topic. This examination of relevant literature serves two purposes: It complements the theoretical definitions laid out in the previous section by contextualizing them with a greater layer of detail, and it introduces the empirical side of this study by examining how previous scholars have estimated credit yield spreads.

4. Literature review

The literature review will be split in two distinctive sections. The first section focuses on empirical studies of bond spreads, starting with a historical breakdown of corporate bond analysis. This will be followed by a review of literature relevant to the topics introduced in the theoretical section, such as liquidity risk and credit risk. The second section is focused on previous studies that have employed textual data in the prediction of bond spreads and serves as an introduction to the study performed in this paper.

4.1 Empirical studies of Corporate Bonds spreads and performance

One of the earliest empirical investigations of the corporate bond market was conducted in 1959 by Lawrence Fisher. Fisher (1959) investigated the determinants of the risk premia found in corporate bonds. In 1968, Altman developed a framework for determining the risk of default, a main component in the risk premium, through the analytical technique ratio analysis with the development of his Z score (Altman, 1968). The pricing of bonds, with emphasis on the risky nature of the asset, was further cemented through the development of the structural approach to pricing risky debt developed by Merton (1974). Merton Applied the option-pricing model developed by Black and Scholes (1973) to pricing risky debt. Merton argued that creditors are long the assets of the company in question and short a put option on the assets of the company with a strike value equal to the face value of the debt. He then argued how the spread of a risky bond is equal to the value of the put-option, which can be calculated using the Black-Scholes formula. According to the formula, the spread should be a function of a firms leverage, the maturity of the debt instrument, the volatility of the company's assets and the risk-free rate (Merton, 1974). Since then, many studies have tried to use the structural approach with a focus on credit risk when studying the pricing of bonds (Black & Cox, 1976; Sundaresan et al, 1993; Shimko et al, 1993; Nielsen et al, 1993; Longstaff & Schwartz, 1995; Anderson & Sundaresan, 1996; Jarrow & Turnbull, 1995; Lando, 1998; Collin-Duffresne & Goldstein, 2001; Campbell & Taskler, 2003; Butera & Faff, 2006; Ericsson et al, 2009). However, studies like Huang & Huang

(2003), Huang & Huang (2012), Longstaff et al (2005), Ericsson & Renault (2006), and Dick-Nielsen et al (2012), document that the structural approach and the models employed by it underestimate yield risk, implying that the yield spread of corporate bonds contains other premia besides the credit risk premium. This gap between the observed spreads and the spread required to compensate for expected default losses has been dubbed '*the credit risk puzzle'* (Amato & Remolona, 2003). Following the finding that structural models and studies focusing only on the risk premium required to compensate for expected default losses, a strand literature emerged focusing on the significance of liquidity for explaining the yield spreads observed in corporate bond markets. As a consequence, most modern studies either focus on explaining the credit risk or the liquidity premium while controlling for the other. We first investigate studies examining liquidity risk, before turning the focus to credit risk.

4.1.1 Literature on the Liquidity risk premium:

Liquidity has historically been difficult to measure directly before transactional data was more readily available to researchers. Early studies of the importance of the liquidity premium relies on liquidity proxies instead of direct liquidity measurements. Longstaff et al (2005) used the Credit Default Swap (CDS) market for corporate debt to obtain direct measures of the size of the default and liquidity components of corporate yield spreads. They find that the default component represents the majority of the spreads. Even for the highest ranked (AAA) bonds in their sample, the default risk premium accounted for more than 50%, a figure that is much higher for high yield ranked companies. But robust evidence for a significant nondefault components of spreads is also found, especially for higher ranked companies. The nondefault component of spreads is strongly related to measures of bond-specific illiquidity (Longstaff et al, 2005). Huang & Huang (2003), uses data from the US bond market to try and determine how big a portion of the yield spread is attributable to credit risk and can be explained through structural models. They conclude that for investment grade bonds, the fraction of the spread attributable to it is relatively low, as credit risk only accounts for around 30% of the spread on the investment grade bonds in their US bonds only sample, while accounting variables significant for credit risk accounts for a much higher degree of the spread on high yield bonds (Huang & Huang 2003). Ericsson & Renault (2006), develop a structural bond valuation model to capture both credit and liquidity risk, with a focus on distressed debt (debt where the obligations is unlikely to be met). They find that the renegotiation of distressed debt is affected by the illiquidity of the distressed debt market. As default becomes increasingly likely the part of the spread attributable to illiquidity increases. They also find evidence of a positive correlation between the illiquidity and default components of yield spreads (Ericsson & Renault, 2006).

Chen et al (2007) uses more direct measures of liquidity. Using a battery of liquidity measures, amongst them the direct bid and ask spreads of 4.000 corporate bonds (both investment grade and high yield) obtained through Bloomberg. They find statistically significant results of illiquid bonds earning higher yield spreads and concludes that liquidity is priced into the spreads of

corporate bonds (Chen et al, 2007). Bao et al (2011), uses transaction data from 2003 to 2009, and finds that the illiquidity in bonds is both substantial and significantly greater than what can be explained through the bid and ask spreads that laid the foundation for previous analysis. They find a strong link between bond prices and illiquidity and find that changes in liquidity in the bond market can explain much of the variation of the corporate bond spreads over time (Bao et al, 2011). Similarly, Lin et al (2011) studies a cross section of bonds from 1994-2009, finding that the return on bonds with high sensitivity to the liquidity in the markets exceed that of bonds with low sensitivity to the market liquidity by around 4% p.a. Thus, it is concluded in accordance with Bao et al (2011) that liquidity risk is an important determinant when calculating expected returns of corporate bonds. Finally, Dick-Nielsen et al (2012) examine the illiquidity component of corporate spread before and after the subprime crisis of 2008. It is found that the spread contribution from illiquidity increases drastically with the onset of the subprime crisis. Additionally, it is shown that this increase is slow and persistent for investment grade bonds while stronger but also more short-lived for high yield bonds (Dick-Nielsen et al, 2012). The crisis proved that bonds become more illiquid in times of financial distress, which in the peak of a crisis will hit the liquidity for lower grade bonds harder, as the high yield markets will dry up as a consequence of a "flight to quality" (migration towards AAA rated products).

While some of the studies in the above section showed how the illiquidity premium holds strong explanatory power for assessing the development of the yield spreads for the aggregate corporate bond markets (Chen et al, 2007; Lin et al, 2011; Bao et al, 2011), it is not the right tool for assessing the difference in performance of individual bonds issued in the same market. The aim of this paper is to develop both a theoretical framework and empirical model that can be used by asset managers in the high yield market to predict winners and losers amongst individual bonds issued within the same market and timeframe. For this task, liquidity is not the right parameter of analysis (Fridson, 2018). Furthermore, while the literature on the liquidity premium shows the overall share of the yield spread attributable to illiquidity is a significant part for corporate bonds in general, researchers have found this share to be a minor and insignificant part for high yield bonds, due to the much larger risk spreads found in this asset class (Dick-Nielsen et al, 2012; Bao et al, 2011; Longstaff et al, 2005; Chen et al, 2007). Therefore, the literature on credit risk is better suited and more directly relevant for developing an analytical framework and parameters for the objective of this paper. However, the importance of illiquidity on the overall bond yield, means that it is a parameter that should ideally be controlled for in a model trying to predict spreads.

4.1.2 Literature on credit risk premium and default risk:

As mentioned, Fisher (1959) developed the earliest empirical assessments of the credit risk spread for corporate bonds. Fisher investigated the US bond market and developed the following four hypotheses: 1) that the average risk premium of a firm's bonds depends first on the default risk and second on their marketability. 2) the risk of default can be estimated by three variables: '*the coefficient of variation in firm's Net Income for the last nine years*' (stability of the profitability),

'*the length of time the firm has been operating without forcing its creditors to take a loss*' (consistent managerial performance), and '*the ratio of the market value of the equity in the firm to the par value of the firms debt*' (the capital structure / leverage). 3) the 'marketability' of a firm can be estimated using a single variable: the market value of all the publicly traded bonds the firm has outstanding (an early alternative to liquidity). 4) The logarithm of the average risk premium can be estimated by a linear function of the logarithms of the four variables just listed. Fisher tested this on a sample of 366 observations of bonds of industrial companies over a 5-year timespan and was able to explain 74 percent of the variations in risk premia.

Altman (1968) is one of the first to create a quantifiable framework using ratio analysis as an analytical technique to predict the probability of default for corporate bonds of publicly traded companies. This was done through the development of the Z-score, where a set of financial ratios was combined in a discriminant analysis approach to the problem of predicting corporate bond defaults (Altman, 1968). The Z-score is calculated as follows:

$$Z = .012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + .999X_5$$

Where, X_1 = Working Capital/Total Assets, X_2 = Retained Earnings/Total Assets, X_3 = EBIT/Total Assets, X_4 = Market Value of Equity / Book Value of Debt, X_5 = Sales / Total Assets.

The discriminant analysis of the Z-score proved to be highly accurate on Altman's sample, predicting 94 percent of bankruptcies in the original sample correctly (Altman, 1968). Along the same lines a study by Beaver (1967) concluded that cash flow to debt ratio was the best single ratio predictor. The Z-score framework was revisited again by Altman in 2000 and adjusted to a Z''-score that also works for private companies, as the market value of equity was changed for the book value of equity. He found the new Z''-score to be slightly less accurate than the original Z-score (Altman, 2000).

Other studies have built upon the framework of ratio analysis in attempts to predict the performance of corporate bonds. Khurana & Raman (2003) examines the relevance of long-term fundamentals for default risk, as they setup a regression model using a long range of fundamental performance indicators including Altman's Z-score, to predict YTMs for new issues of corporate bonds. It was found that both the aggregate fundamental score and the individual fundamentals provide incremental explanatory power in pricing new bonds, as indicators of expected future earnings and solvency, and that they hold significant explanatory power above that of the published bond rating classifications (Kuhrana & Raman, 2003). Ohlson (1980) proposed a logit model of nine explanatory variables in his O-score model. This was later built upon by Campbell et al (2008) who combined both current and lagged accounting and market information in a single logit model of eight variables (the C-score) to predict defaults. Castagnolo & Ferro (2014) assessed and compared the forecast ability of a range of credit risk models, to test whether they can accurately predict default events, and found the O-score of Ohlson (1980) to outperform the

others, amongst them Altman's Z-score, the option valuation model of Merton (1976) and the C score of Campbell et al (2008). Butera & Faff (2006), seeks to convert historical probability of defaults into a forward looking model using accounting and economic fundamentals to predict default that also works for private companies using a sample of private Italian companies (Butera & Faff, 2006). The fundamentals of analysis varies from study to study but a general consensus in the literature is that proper fundamental ratio analysis should include parameters across five standard categories: 1) liquidity (of the company, not the market), 2) profitability, 3) leverage, 4) solvency, and 5) activity (Altman, 2000). Huffman & Ward (1996) applies such framework to the high yield market, in an attempt to predict default of high yield bonds based upon public information at the time of issuance, using multivariate analysis through logistic regression. They employ variables from the 5 standard categories as well as arguing that measures of future growth and managerial slack is necessary additions for high yield bonds (Huffman & Ward, 1996). The latter variable is measured using the ratio of the capital structure that comes from debt, due to the cash constraints that debt places on management (Hart, 2001; Bruner, 1988).

Other papers make more broad assessments of credit risk in the corporate bond market. Elton et al (2001) attempts to explain the difference between the spread of corporate bonds and on government treasuries. They expand the literature by also showing that a substantial portion of the spread is attributable to taxes. This is because, in some states in the US, interest payments on corporate bonds are taxable whereas interest payments on government bonds are not. They find the remainder of the spread to be attributable to credit risk (Elton et al, 2001). Collin-Dufresne & Goldstein (2001) use a structural approach to investigate the relationship between credit ratios and leverage ratios. In their paper, they create a framework where firms adjust their capital structure to reflect changes in asset value. The framework generates stationary (mean-reverting and non time-dependent) leverage ratios for the sample companies and finds that the term structure of credit spreads of speculative debt should in general be upward-sloping (Collin-Dufresne & Goldstein, 2001). Campbell & Taskler (2003) shows that idiosyncratic firm-level volatility, estimated by equity volatility, can explain as much of the variation in credit spreads as credit ratings can. Ericsson et al (2009) tests the significance of a range of theoretical factors determining the credit yield spread in many structural models using the CDS rate as a proxy for the credit risk premium, and find that leverage and firm-level equity volatility are the two most significant explanators for variations in the credit yield spread (Ericsson et al, 2009). The explanatory power, captured the firm-level volatility measured in both Ericsson et al (2009) and Campbell & Taskler (2003) by the volatility of the equity, can be seen as an estimate of the underlying risk of the company as a whole and its future profitability. This information could potentially be better captured through the text in the risk section of the bond prospectuses which can be more systematically analyzed through modern machine learning and NLP techniques. But before we leave the world of fixed income to investigate literature that focuses on this for equity

prospectuses, it is worth examining studies that focus exclusively on high yield corporate bonds in greater detail.

4.1.3 Empirical studies with specific focus on high yield corporate bonds:

Many of the aforementioned studies are carried out on a sample of investment grade bonds or on a sample of corporate bonds as a whole, which implies that a majority of the sample is investment grade bonds. As mentioned, the higher risk profile of high yield bonds means that information about the future performance of the company and estimations of the probability of default are more important for spreads. This combined with the relative smaller size of the market and lower liquidity compared to investment grade bonds (AFME, 2020), means that literature specifically on high yield bond market is relevant to strengthen the understanding of the effect of these differences. However, the literature on high yield bonds is much sparser than the relative rich literature on corporate bonds in general. Especially literature on the European high yield bond market is limited, due to its relatively young age and the scarcity of data availability, as many of the European HY issuers are privately held companies.

Because of the high risk-profile of non-investment-grade bonds, offerings of HY debt has historically been labeled as 'story bonds' for which quantitative, objective valuation criteria for pricing or spread determination are difficult to establish. Analysts would have to learn about the nature of these companies using more in-depth ad hoc analysis (Fridson & Garman, 1998). Fridson & Garman (1998) succeeds in explaining 56% of the variance in risk premiums of US High Yield bonds through quantifiable factors, such as ratings, term structure, and secondary market spreads. The high-risk profile also makes prediction of defaults probabilities of even higher value, and several studies such as Huffman & Ward (1996) have applied different structural approaches and multivariate analysis to predict default rates in high yield bonds. Huffman & Ward (1996) was able to correctly predict 73,3 percent of the defaulted bonds and 68,6 percent of the non-defaulted bonds. Applying machine learning to the prediction of high yield bond performance is not a completely unexplored avenue. Ashby and Kumar (1996), tried to apply early stage of neural network models using one hidden layer, achieving an 89% accuracy in classification of 56 records.

Fridson (1990) found that initial pricing holds explanatory power as a predictor for subsequent performance of HY Bonds. He found that high yield bonds that eventually run into serious credit problems, generally do not enter the market as an average quality issue. Instead, the spread that they go to market within initial offerings tag them as potential distressed candidates. As a consequence, just going for the maximum yield within a class of rating appears to be a questionable investment strategy and superior analytical skills are needed (Fridson, 1990). In a 1996 paper, Fridson and Gao (1996) investigates systematic underpricing in the initial offering of high yield bonds. It is found that primary issues provide superior risk-adjusted return in the period subsequent to the issuance, and the model is able to explain 64% of the variance in the

primary vs secondary yield spread, using proxies for supply, demand and liquidity in the high yield market.

In more recent studies Gentry et al (2010) investigates the statistical properties of the credit risk spread for high yield bonds and to analyze the influence of a set of variables that are expected to have an effect on the credit risk spread. They study a group of 9 variables through multivariate analysis and finds that the strongest impact comes from a range of default risk variables. Similar to the findings in the general corporate bond markets by Campbell & Taskler (2003) and Ericsson et al (2009) on the relationship between bond spreads and the volatility of the underlying stock, Wu & Zhang (2014) examines the relationship between the return of the high yield bond market and the stock market. The study finds that stocks lead high yield return and that this lead-lag relationship is strongest during bear markets since a downward trend in the stock market implies a high likelihood of the exercise of the short position equity put embedded in a high yield bond at maturity, similar to the model developed by Merton (1974). Li et al (2014) investigate the performance of high yield bonds compared to investment grade. They show that, when a normal distribution is assumed, high yield bonds achieve higher expected return *and* lower standard deviation. But also, that both high yield bonds and investment grade bonds exhibits fat tails, which means that distributions that allows for tails or skewness should be applied when investigating high yield bond risk profiles.

Not all high yield debt of equal credit rating has the same risk profile or demand the same credit risk spread. Fridson et al (2016) shows that yield spreads for B rated corporate bonds were greater on senior ranked than on subordinated ranked debt. John et al (2010) also found that presumed less risky senior bonds had wider spreads than the like-rated subordinated issues. This indicates that the lower probability of default of a subordinated B rated issue[4] could be expected to more than offset increased loss given default of a senior ranked B1 issuer, resulting in a lower total expected loss for the subordinated issue (Fridson et al, 2016).

To summarize, literature on corporate bonds in general and high yield bonds in particular, highlights the importance of correctly estimating the credit risk when attempting to explain overall yield spreads. With high yield bonds being the type of asset with least available information while being most affected by credit risk, increasing the amount of data used to predict performance by including qualitative parameters from bond prospectuses seems like an appealing preposition. This is also well aligned with the idea of HY bonds as “story bonds” in need of additional credit analysis (Fridson & Garman, 1998), and allows for better prediction of subsequent performance as the information upon issue increases (Fridson, 1990)

[4] The probability of default is lower, as the comparable senior ranked bond issued by the same company would be higher rated, e.g. BB

4.2 Empirical studies on the performance of financial securities using text data

The scope of this section is not to present the theory and literature behind the text processing and machine learning models that will be used for analysis in this paper, as this will be the focus of the subsequent methodology section. Instead, the aim of this section is to lay out the existing literature on the topic of applying Natural Language Processing, which is the intersection between machine learning and linguistics, to analysis of financial securities' performance, and thus lay out both existing best practice frameworks, and to provide further evidence for the validity of using linguistic and textual data as an input in more quantitative models of the financial markets.

The field of using textual analysis as an input for security analysis is relatively young due to the young age of the technology that enables it. But recent advances in NLP and machine learning technology has allowed researchers to make more in depth analysis using text. While there are not any widely known or acknowledged research papers using NLP techniques on data for the high yield bond market, a range of researchers have tried to use NLP analysis to explain trends in the equity markets, which will be the focus of this section.

One of the first inquiries into using NLP analysis on the equity markets are Tetlock (2007). Tetlock (2007) tries to analyze the relationship between the content published by the media and the subsequent movements on the stock market, through a quantitative analysis of daily stock quotes on the S&P 500 index and NLP analysis of a popular wall street journal column. Tetlock (2007) uses the Harvard IV-4 psychosocial dictionary to categories each words of the column into various categories such as 'positive' or 'negative'. Tetlock (2007) then proceeds to use the count of these categories as a proxy for the mood or tone in the media, and finds that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. In Tetlock et al (2008), the unit of analysis is moved from a market level to an individual company level, examining whether a simple quantitative measure of language can be used to predict individual firms' accounting earnings and stock returns. In the analysis, the use a negative word score as an input to their quantitative model, based on the Harvard IV-4 psychosocial dictionary. It is found that the fraction of negative words in firm-specific news stories forecasts low firm earnings (Tetlock et al, 2008).

In a 2011 paper, Loughran & McDonald found that wordlists based on general dictionaries, such as the Harvard IV-4 psychosocial dictionary, often misclassify words in financial texts. In their sample, three fourths of all the words classified as 'negative' by the Harvard dictionary, are typically not considered negative in a financial context. They proceed to develop their own wordlist using a Bag-of-Words method and weighing the importance of words using a TF-IDF matrix, which can then be used as input to train a classifier with machine learning techniques[5]. They combine this wordlist with 5 other wordlists to analyze the relationship between the tone

[5] Both Bag-of-Words and TF-IDF will be explained in greater detail later in the paper

in company annual reports and the trading-volume, volatility, return, and risk of unexpected earnings decline of the underlying stock and company. Jegadesh & Wu (2013), developed their own weighting scheme as an alternative to the TF-IDF, and proposed an approach that assigns weights for each word based on market reactions to documents containing those words. This is a manual approach that is quite similar to the way modern machine learning algorithms would assign weights through brute force of computing power (Géron, 2017).

In the context of NLP analysis, company initial public offerings share some similarities with a high yield bond offering, since like a bond offering, all the information about the IPO and the company of issuance is presented in a long prospectus. Several researchers have used linguistic and textual analysis of IPO prospectuses to shed new light on the performance of IPOs. Loughran & McDonald (2011) use tone analysis on form S-1 filings (the first to the SEC[6] in an IPO), and finds that IPOs with high levels of uncertain text in their S-1 filings have higher first-day returns, absolute offer price revisions, and subsequent volatility. Bartov (2011) uses NLP to analyze the relationship between the content of the risk section of the IPO prospectuses and future earnings and analyst forecast. They employ three different approaches to quantify the earnings downside to textual risk information, all using a wordlist developed through reading numerous prospectuses. They find that the textual downside earnings risk information in the risk section of the IPO prospectuses are correlated with future earnings and analysts' forecast error, but not with the analyst forecast themselves, which also proves that the text of the risk section can provide explanatory power beyond what is available through analyst predictions (Bartov, 2011). Fisher et al (2015) tries to explain the first day returns on IPOs through NLP and Past-of-Speech (PoS) analysis of the risk section of IPO prospectuses. They find that sentiment wordlists are significantly correlated with first day returns, but also find that words that are underrepresented on such lists have even more explanatory power. Furthermore, the PoS analysis shows that nouns and adjectives carry the strongest explanatory power (Fisher et al, 2015). In a more recent paper, Yan et al (2019) show that the negative or uncertain tone in prospectuses lowers the stocks' long-term returns, in an analysis of the relationship between IPO tone and initial IPO returns on the Chinese market.

Hanley & Hoberg (2010) uses the textual content of IPO prospectuses as an indicator for how much time the underwriter has spent on making the premarket due diligence. If the content is very similar to general content (close to the mean), a lot of the writing is generic writing that have been taken from previous issues, which signal low time spent on due diligence. More unique IPO texts signals more premarket effort by the underwriter. Deokar & Tao (2015) uses NLP analysis on IPO prospectuses on a sentence level and develop the FOCAS-IE framework for analyzing financial text documents (Feature-Oriented, Context-Aware, Systematic Information Extraction). They combine the text sentiment derived from the FOCAS-IE framework with several modern predictive machine learning methods and find that pricing can be better predicted using the text

[6] Security and Exchange Commission

features alongside the quantitative IPO features than with the quantitative IPO features alone. They achieve the best result with a Decision tree model, but also significantly outperforms the quantitative only model with an artificial neural network model (Deokar & Tao, 2015). These findings are built upon in a paper by Deokar et al (2018). In this paper, they develop a forward-looking-statement extractor and classifier using deep learning architecture that is found to outperform any prior model. Forward-looking-statements are used to analyze pre-IPO price revisions and post-IPO first day returns, finding that FLS features are more predictive for pre-IPO as compared to post-IPO valuation prediction (Deokar et al, 2018).

NLP analysis has also been applied on other content that IPO prospectuses to explain stock returns. Several researchers have used twitter text data as a proxy for public mood and sentiment, to predict movements in the stock market (Patel, 2016; Fabozzi et al, 2016). Patel (2016) uses the 'TextBlob'[7] sentiment classifier to extract sentiment from twitter data for examining the relationship between twitter sentiment and stock market movements. He uses three machine learning techniques: Naïve Bayes, Decision Tree, and Support Vector Machines, and find that the model can successfully predict stock movements based on twitter sentiment. Fabozzi et al (2016) conducts a similar study but uses a TF-IDF matrix of text content rather than sentiment to try to predict the stock movements.

Generally, the literature on text analysis of financial material comes down to three different steps: 1) content labeling, 2) classifier input generation, 3) classification/regression (Emadzadeh et al, 2010). Various researchers have applied different methods for each of these steps, depending on the nature of their analysis and the data available to them. Content labeling can be done in two ways, manually and automated. Manually, experts can label the content (e.g. relevant vs. not relevant, positive vs. negative etc.) and these labels can be used as input for prediction. In automated labeling the text data is combined with other data used for labeling, such as price movements in the underlying security (Emadzadeh et al, 2010). In classifier input generation, the two important choices are feature selection and feature weighting[8]. Most commonly used in the literature is the method of creating a term dictionary, often in the form of a Bag-of-Words (Deokar et al, 2018). But some researchers have had success using concept maps or topic modelling as the features using Latent Diriclecht Allocation (LDA) (Dey et al, 2008; Deokar et al, 2018). For term weighting a few early papers have used binary BoW weighting (Knolmeyer & Mittermayer, 2006, Halgamuge et al, 2007), but the most common and popular weighting method is the TF-IDF (Fung et al, 2002; Alberg et al, 2007; Chen & Schumaker, 2009; Loughran & McDonald, 2011). For classification, most of the authors of earlier papers have used Support Vector Machine (SVM) as their classification algorithm (Emadzadeh et al, 2010). But in later papers, researchers have

[7] TextBlob is a Python NLP library. See section on text analysis for further explanation.
[8] A feature, in the context of text analysis and machine learning, is another word for variable.

achieved high success with decision tree models, ensembles of different kinds and neural network models (Deokar et al, 2018).

From review of the literature it is clear that NLP analysis is relevant and can be effective when attempting to predict the performance of financial instruments. But with no previous analysis done on high yield bonds, this study will attempt to contribute to the literature by demonstrating the relevance of text analysis for this asset class. Additionally, as the field of NLP is in rapid and constant development, it will be possible to apply models more sophisticated than those employed by early NLP scholars, thereby adding to the understanding of how text in bond prospectuses affect the pricing and subsequent performance of bonds.

5. Theoretical framework for choice of variables

Building on the theory presented, as well as the empirical literature review, the purpose of this section is to explain the theoretical framework behind our empirical model. The section will cover theoretical considerations when choosing independent and dependent variables for examining high yield bond performance. An outline of the best theoretical predictors will be provided, along with how they can be proxied by measurable variables.

The choice of variables can be divided into two groups: the choice of dependent variable, the variable the model aims to explain and/or predict; and the choices of independent variables that should have a theoretical effect or correlation with the dependent variable. The choice of dependent variable in this paper concerns the choice of the right measure for bond performance. The choice of independent variables falls into four overall categories. 1) accounting variables related to the issuing firm, that aims at capturing the current and expected future performance of the issuing firm. 2) other non-accounting firm specific variables. 3) issue specific variables. And 4) choice of text input.

5.1 Choice of dependent variable

The first step to building an empirical model is to choose the variable to measure. This variable will function as label for input data and is the variable to be explained or predicted. It therefore needs to fit the research question the paper is trying to answer. In this paper we aim to assess high yield bond spreads. When choosing a variable to measure the performance of financial instruments, an important distinction is between the choice of an ex-ante or an ex-post variable. An ex-post variable would be historical performance of the security, and whether it has outperformed or underperformed the overall market. Classic examples of this are alpha analysis, where the researcher examines whether any security has given higher or lower realized returns than the risk-profile of the investment would require. Example of this in the high yield space is Trainor's (2010) analysis of HY-fund managers' performance. Ex-ante measures for bonds are the yield-measures, which is the compensation investors require to invest at any given day, and thus include the expectations for the future. This paper aims to explain the spreads at the time of issue,

and at any given date thereafter, and therefore apply an ex-ante measure. Furthermore, it is not the objective to explain the part of the yield of high yield bonds determined by the risk-free rate, but the spread that they provide above risk-free rate. This exclude choosing yield to worst as the dependent variable. As such, the first step is to choose the right spread measurement. The three commonly used spreads are T-spread, G-spread, and Z-spread (Fabozzi, 2013).

T-spread

The T-spread is simply the yield spread that the bond trades above treasury benchmark bond, which will typically be a treasury bond issued by the government of the official country of the company or the government of the biggest market for the company.

$$T_{spread} = Yield_{bond} - Yield_{Treasury}$$

It is important that the benchmark treasury bond is denominated in the same currency as the bond of analysis, so the spread only captures risk and not any differences in expected inflation rates of the two currencies. However, this measure does not consider the underlying yield curve of the treasuries

G-spread

Government spread, known as G-spread, is similar to T-spreads, but takes into account the underlying yield curve of the treasuries. It is the spread over the exact interpolated point of the benchmark treasury yield curve:

$$G_{spread} = Yield_{bond} - Yield_{Treasury\ of\ same\ maturity\ horizon}$$

Z-spread

Zero volatility spread, known as Z-spread is the spread that must be added to each spot interest rate along the yield curve to make the value of the cash flows equal to the price of the bonds. The Z-spread is calculated as:

$$P_{bond} = \frac{CF_1}{(1 + S_1 + Z)} + \frac{CF_2}{(1 + S_2 + Z)} \cdots \frac{CF_n}{(1 + S_n + Z)}$$

Where P_{bond} is the price of the bond CF_i is the respective cashflows of the bonds and S_i is the spot rate on the government yield curve *i* periods into the future.

As the Z-spread best captures the full time-aspect of the bonds and therefore leaves only risk in the spread, it would in theory be the best choice of spread variable. However, sometimes reliable Z-spreads can be difficult to obtain so in the case of data limitations, the G-spread will function as a good substitute, as this also captures most of the time aspect of the bond, and is often very similar to the Z-spread (Fabozzi, 2013). As the proper yield measure of the bond, from which the treasury yields will be subtracted, we use Yield-To-Worst. As explained in the theory section, call

features on the bonds will sometime lead to lower yield to the call date and price than to maturity, and as rational companies can be expected to call the bond in such a case, the YTW will be a more appropriate yield measure than the YTM. It should be noted that G-spread is the most widely applied measure for the yields spread of corporate bonds (Longstaff et al, 2005; Fridson, 1998; Gentry et al, 2010).

5.2 Choice of independent variables

5.2.1 Accounting variables

The following section will describe our choice of accounting variables for independent variables for analysis in our model. As laid out in the theory section, the spread premia paid on high yield bonds are paid to compensate for the expected loss on the bonds, given as:

$$Expected\ Loss = probability\ of\ default * loss\ given\ default(100\% - recovery\ rate)$$

The variables are thus divided into variables chosen as estimators of the probability of default and estimators of loss given default. Some variables will have explanatory power for both, but for clarity and to provide structure, they are divided into the two subgroups in the following section.

5.2.2 Probability of default variables

As previously mention, default occurs when a borrower fails to meet the obligations laid out in the debt contracts. When analyzing high yield bonds either as an analyst or through quantitative models, one needs to look at metrics that can be used for calculating the risk of this happening (Fridson, 2018).

5.2.2.2 Capital Structure / Equity Cushion

One of the first key metrics to analyze is the capital structure of the company – which proportion of the full value of the company does the debt account for. The capital that equity holders provide to the company is subordinated in the capital structure to any form of debt.

This means that shareholders will take the loses before any value is lost for debtholders. As a result, in general, the company should be able to avoid default as long as value remains in the equity tranche of the company, as the company can use that capital to service the debt, and in the case of a technical default, the recovery rate for the bond holders should be 100% if any value remains for the equity holders. For that reason, the level of equity to the total value of the company is often referred to as the equity cushion, as it functions as a cushion of equity that will have to be wiped out before any value is lost for the bond holders. Ideally, one would use the market value of the company and the equity of the company:

$$\boldsymbol{Equity\ Cushion} = \frac{Equity}{Enterprise\ value} = \frac{EV - Net\ Debt}{EV}$$

However, for private companies such data is not available. Instead of the market value of the company and the equity, the book value of the company, given by the book value of Total Assets from the balance sheet can be used as another metric:

$$\boldsymbol{Share\ of\ debt} = \frac{Net\ Debt}{Total\ Assets}$$

Higher equity cushion and lower share of debt will theoretically lead to both a lower risk of default and a lower loss given default and would be expected to be correlated with the spread a bond is trading at.

5.2.2.3 Leverage

One of the most widely applied metrics is the level of debt a company holds relative to its profitability, known as leverage (Fridson, 2018). By holding the debt level relative to the earnings of a company, it provides a sense of how sustainable the level of debt is, or in other terms, what the risk of default on the debt is. The most widely used leverage metric is Debt / EBITDA:

$$\boldsymbol{Leverage} = \frac{Net\ Debt}{EBITDA}$$

As this metric is available for both private and public companies (given you have access to the financials of the private companies). Furthermore, EBITDA is used as the profitability metric for three reasons: Firstly, it is the earnings measure that is closest related to the actual cash generated from the earnings, as it is calculated before depreciations and amortization. Secondly, it is the earnings available to service the debt as taxes are first calculated after the debt is serviced. Thirdly, it estimates cash generated while cutting away noise such as one-off expenses, write-offs etc. The level of leverage can change in two ways, either companies increase or decrease their level of debt to a desired leverage level, or the EBITDA improves or deteriorates while the debt level is stable. The sustainable level of leverage varies greatly from industry to industry, depending on how secure or stable the earnings in the industry are. However, all else equal, a higher leverage would result in an increased probability of default. Leverage metrics in the form of Net debt to EBITDA or as a share of the full capital structure is used in many research papers

on the credit risk spreads of bonds such as Fisher (1959), Blume et al (1998), Collin-Duffresne et al (2001), Ericsson et al (2009), Kovner & Wei (2012), Ashby & Kumar (1996), who all found leverage terms to be an important indicator for explaining credit risk spreads of bonds. Similarly, it is a key component in both Altman's Z-score (Altman, 1968), Campbell's C-score (Campbell et al, 2008) and in Moody's Risk Calculator 4.0 (Dwyer et al, 2014).

5.2.2.4 Debt Coverage & Cash Generation

While the leverage can give an indication whether the overall level of debt is sustainable in general, it is also important to analyze the ability of the company to meet interest payment requirements in the short term. Even though the company may earn enough to service the obligation on its debt in the long term, if it fails to generate the liquidity to meet one of its interest payments on a short-term basis, it will go into default. A company need to turn earning into cash flow that can be used to service the debt, on a basis stable enough to meet every interest payment. A widely used metric for debt coverage is the Interest Coverage Ratio (Ashby & Kumar, 1996). Interest Coverage Ratio can be defined slightly different, some define it as EBIT/Net Interest payments, others as EBITDA-CAPEX / Net Interest payments. In this paper we adopt the common definition (Fridson, 2018):

$$\boldsymbol{Interest\ Coverage\ Ratio} = \frac{EBITDA}{Net\ Interest\ Expense}$$

As EBITDA is a metric that allows for less adjusting of accruals by management, and is more stable from year to year, than measures including CAPEX.

But as mentioned, earnings do not equal cash flow, as companies can have many non-cash flow earnings or cost and many non-earnings-related cash flows (Berk & DeMarzo, 2017). Therefore, a measure for cash generation is needed. As long as the debt bears normal coupons and not Payment-In-Kind coupons then, holding the earnings level constant, a higher cash flow generation would lead to a lower risk of default (Fabozzi, 2013). As with the other metrics, the total cash flows will vary with the size of the company, so it should be kept relative the level of debt:

$$\boldsymbol{Cash\ Generation} = \frac{FCF\ before\ financing}{Net\ Debt}$$

$$\boldsymbol{Cash\ Generation\ from\ operations} = \frac{Operating\ Cash\ Flow}{Net\ Debt}$$

Cash generation is an indication for the firms more direct ability to generate cash, while cash generation from operations is a metric for the firms more core cash generating ability, as the cash generated from operations will have less noise and one-off instances.

Similarly, to the ability for the company's ability to generate cash, metrics to estimate the current liquidity position of the company is applied. The current liquidity level can be an indicator of the level of cash the company are able to hold in reserve as a cushion towards unexpected dips in earnings or other increases in cash requirements, which should affect the chance of default negatively. On the other hand, to high levels of cash, could also be an indicator that the company is run inefficiently (Berk & DeMarzo, 2017), which is a negative indicator for long term performance, which leaves the theoretical effect on bonds with long term maturity more ambiguous. We apply two widely used liquidity measures (Altman, 1968; Ohlson, 1980; Dwyer et al, 2014; Khurana & Raman, 2003):

$$\boldsymbol{Cash\ Ratio} = \frac{Cash\ \&\ cash\ like\ items}{Total\ Assets}$$

$$\boldsymbol{Working\ Capital\ to\ Total\ ASsets} = \frac{Working\ Capital}{Total\ Assets}$$

However, due to the data limitation, for our dataset for private companies as a proxy we apply:

$$\boldsymbol{Cash\ Ratio} = \frac{Cash\ \&\ cash\ like\ items}{Sales}$$

5.2.2.5 Profitability

Analyzing whether the company can meet its coupon/interest payments is not enough on its own. Eventually the company will also need to pay back the bond or refinance. Thus, the analysis requires a metric for profitability that can be an indicator on the more long-term financial health of the company (Berk & DeMarzo, 2017). Furthermore, profitability may be an indicator on the degree of competition in the industry the company operate in. The most used metrics for profitability is earnings relative to sales (margin) or assets (Return on Assets) (Ohlson, ??; Campbell et al, 2008, Altman, 1968). As previously mentioned, the most relevant earnings metric for debt analysis is EBITDA, and accordingly we use the following profitability metrics in the paper:

$$\boldsymbol{Adjusted\ EBITDA\ Margin} = \frac{Adjusted\ EBITDA}{Sales}$$

$$\boldsymbol{RoA} = \frac{Adjusted\ EBITDA}{Total\ Assets}$$

Additionally, in line with Khurana & Raman (2003) we also employ

$$\boldsymbol{NI\ margin} = \frac{Net\ Income}{Sales}$$

Higher profitability will make it easier for the company to meet the debt obligations, and thus can be expected to be negatively correlated with the risk of default, all else equal.

5.2.2.6 Growth

Payments on debt falls in the future, thus current level of key metrics are only relevant indicators of the ability to pay back the debt, as far as they reflect future levels. Consequently, one should take the growth of the company into account as well. Because of the dynamics of economies of scale and scope, in most cases, high growth will eventually lead to increased profits and thus ability to service the debt. As indicator for growth we look at the compounded annual growth rate (CAGR) for the last three year for EBITDA and Sales for the companies.

$$\boldsymbol{Growth_{Sales}} = \left(\frac{Sales_t}{Sales_{t-3}}\right)^{\frac{1}{3}}$$

$$\boldsymbol{Growth_{EBITDA}} = \left(\frac{EBITDA_t}{EBITDA_{t-3}}\right)^{\frac{1}{3}}$$

Three-year average compounded growth rates are used as they regularize the growth levels more than just using a single year's growth e.g. last year's growth.

5.2.3 Loss given default

Predicting the risk of default is not enough to predict the spreads that high yield bonds trade at. Firms with a recovery rate close to a 100% will leave little losses for the debt holders in the case of default and will consequently be able to trade at low spreads even in the presence of high default chances. While the true loss in case of default is unknown until it is too late, it can be

estimated through certain metrics. Thus, the quantitative model should include a metric for the loss given default rate. The recovery rate for the specific bond highly depends on its seniority, but estimates can also be done on an issuer level.

5.2.3.1 Tangibility of assets

Two companies with identical capital structures can have very different recovery rates in a case of default. One reason is that the nature of the assets can vary greatly across firms. In the case of a default on debt, the bondholders will take control of the company and typically try to sell its assets to recoup as much of the value of the bonds as possible. Certain assets are much easier to sell at a fair value in case of a default compared to other. How easy an asset is to sell off depends on how generalizable and transferable it is. In crude terms that means how easily the assets are put to use at another company. Berk & DeMarzo, 2017). While there is no perfect way of measuring how easily sold the assets of a company are, they can be estimated through the tangibility of the assets:

$$\boldsymbol{Tangibility\ of\ Assets} = \frac{Tangible\ Assets}{Total\ Assets}$$

Tangible assets are defined as physical assets, whereas intangible assets are non-physical assets e.g. intellectual property. Tangibility of Assets works as indicator of the expected recovery rate and should therefore be negatively correlated with the spread a high yield bond trades at, all else equal.

The above-mentioned variables combined should give a good indication of both the probability of default and the expected recovery rate for the companies in the sample. Furthermore, the accounting variable chosen for our framework is chosen in line with Altman's (2000) five areas that should be covered when estimating the risk-profile of a company in a quantitative model: 1) liquidity (of the company, not the market), 2) profitability, 3) leverage, 4) solvency, and 5) Activity. Liquidity is captured in the cash ratio and WC / TA. Profitability is captured in the ROA, the NI margin and the Adj EBITDA margin. Leverage is captured in the Adj EBITDA / Net Debt. Solvency is captured through the Interest Coverage ratio, and activity is captured in the Industry dummy variables, which is explained further in the next section.

5.2.4 Issue Specific Variables

The above metrics are firm wide and are estimators of the performance of the issuing company. A part of the spread a bond trades at or is issued at, however, depends on metrics specific to the

underlying bond issue. Any quantitative model would have to take these into account alongside the accounting variables.

5.2.4.1 Coupon payment

The part of the overall risk stemming from coupon payments shift the time-horizon-structure of the interest payments and may affect the overall yield spread required by investors on the bond. Higher coupons would mean a bigger part of the yield gained on the bond falls closer to the existing date, compared to yield gained from a difference in price and principal.

5.2.4.2 Call feature and time to call

As laid out in the theory section, a call feature requires a premium in the yield paid to investors. Consequently, the model needs to control for whether the bond has a call feature, the call price and time to the next call date.

5.2.4.3 Payment-In-Kind Feature

Similarly, the model should control for whether the bond has a PIK interest option, as this would also require a spread premium.

5.2.4.4 Credit Rating

As explained, credit ratings hold a large amount of information about the credit risk of a specific issue. The risk of including this parameter in a model is the risk of perfect multicollinearity between the ratings and some of the other metrics, as they may be key indicators for the assignment of the rating itself. This will needed to be tested for before including credit ratings in any model. In the absence of perfect multicollinearity, ratings are a good metric to include. Furthermore, the rating will function as a benchmark for any data-based model, as it will have to outperform a model that only have the credit ratings as input to add new explanatory power.

5.2.5 Other variables

A few other variables that neither falls in the category of being an accounting variable or are issue specific are included in the model, as they hold potential as explanatory variables or variables that needed to be controlled for.

5.2.5.1 Industry group

The risk of companies in different industries may vary greatly. For instance, utility providers have very secure and stable cash flows, whereas the future cash flows of fashion companies will be much more uncertain. Thus, two companies with identical financials, but where one is a utility provider and the other is a fashion company, will have very different risk profiles. As a consequence, the bond issued by the fashion company will require a much higher spread. As a result, an industry or activity classification will be a potential explanatory variable for predicting bond spreads.

5.2.5.2 Country

The economic and legal conditions may vary from country to country. As with the industry, two companies with identical financials but located in different countries may have different risk profiles. Thus, the country can impact the spread of the bonds, because an investor would prefer the bond from a country with a strong economy and legal framework, rather than one with a more troubled economy and legal framework, if the two bonds are otherwise identical. For example, Lie & Nielsen (2015) found a premium on bonds issued in Southern Europe compared to Northern Europe.

5.2.5.3 Liquidity

As laid out in the literature review, many scholars found that a significant part of the spreads of corporate bonds comes from the lack of liquidity in the corporate bond market (Longstaff et al, 2005; Huang & Huang, 2003; Ericsson & Renault, 2006; Bao et al, 2011). And while they found, the part of the spread deriving from the market liquidity to be low for high yield bonds, due to the high credit risk of those bonds, it is still something that should ideally be incorporated into any model predicting yield spread of high yield bonds. While there is no perfect way to measure the liquidity in the market, there are several ways to estimate it. One approach could be to assume that the credit default swap rate for a firm responds to the credit risk of a firm and estimate the liquidity as a residual. We had originally planned to collect the CDS curves for the bonds in our sample from the trading desk at Capital Four Management, but with the outbreak of the Covid-19 virus, we did not have access to the Capital Four trading desk and were thus not able to obtain data on the CDS curves. Another approach would be to take the spread between the bid and the ask prices. We had originally planned to pull those spreads and calculate the liquidity input as:

$$average\ prices\ for\ the\ day(\frac{Ask - Bid}{Mid})$$

But again because of the Covid-19 epidemic, we were unable to access both the Bloomberg terminal at the Capital Four Management office, and the Bloomberg Terminal at Copenhagen Business School before we managed to collect this data. Consequently, we followed the Covid-19 master thesis instructions set out by Copenhagen Business School to carry on with the empirical data already collected. Lastly, the above to are estimates of the overall market liquidity. The size of the issue can be an indicator for the liquidity of the specific issue and will therefore be included as a substitute into the models. That being said, as previous studies clearly have demonstrated, the importance of liquidity risk in the context of high yield bonds is miniscule compared to that of credit risk, meaning that the absence of a market liquidity measure is unlikely to severely alter results (Longstaff et al, 2005; Chen et al, 2007; Bao et al, 2011; Dick-Nielsen et al, 2012).

5.2.6 Prospectus textual information

While all of the above inputs hold strong information about the performance of a company, they still leave out much detail about the risk and opportunity it faces. Ideally, models should not only

consider this quantitative information but also include the plethora of qualitative information available when analyzing a firm's future prospects. One of the important sources of qualitative information is the prospectuses associated with the bond issue. But which parts of the prospectuses should be used as input for analysis? According to Deokar et al (2018), ideally, the input should be forward-looking in nature rather than focusing on the past, as such information would have more relevance for the pricing and performance of the bond. Also, it should be text describing the company itself, or details surrounding the issue, not captured easily captured by numbers. Consequently, sections such as *Management Discussion & Answers (MD&A), Business description*, *industry description, and* the *Risk Section* would be ideal text inputs for analysis.

5.2.5.1 Management discussion & Answers

The MD&A Section is a section where the management of the company provides commentary on the financial statements of the company, actions planned for the company or how they plan to address certain challenges the company may be facing. All of this information may hold informational value beyond what is captured in the current financial numbers and could provide a good input into the model.

5.2.5.2 Business & Industry description

Are sections where the management describes the dynamics of the industry and the business itself, and important activities of the business. This aims at providing investors with the necessary information about the industry that the business is operating in, to make up their investment decisions. It thus could provide information beyond what is captured by the industry classification variable.

5.2.5.3 Risk section

Risk sections include forward-looking information regarding the potential risks that faces both the company and the bond issue. It both includes risk of significant events that could face future earnings (down-side earnings risk), as well as business risk, legal risk, operating risk, financial risk, political risk, and macro-economic risk. And thus, include relevant forward-looking information regarding the bond and the issuer. Furthermore, because the information covered in the risk section has a clear direction, most often negative, the meaning of terms will be less ambiguous. This reduced ambiguousness of the text makes it ideal as input for machine learning.

With the theoretical underpinnings behind model inputs in place, the paper now moves to the empirical analysis of textual features effect on high yield bond spreads, beginning with a detailed description of the data used for analysis.

6. Data

The study will be drawing on two overall forms of data, both secondary: An independent variable, which consists of data on bond performance, more specifically bond spreads, and a range of dependent variables, including descriptive data on high yield bonds, financial data of the underlying companies as well as textual data in the form of bond prospectuses. The following paragraphs will describe the sources of the data used as well as the data collection process, before presenting descriptive statistics and discussing any steps taken to clean and process the data in an appropriate way for input in analysis. In order to properly cover the different strategies employed for data processing, separate sections will give an in-depth description of the descriptive data for the bonds, accounting data of the underlying firms, textual data from the bond prospectuses and lastly, data on bond spreads.

6.1 Data collection

The study will be conducted using two separate datasets, which are both a collection of information on high yield bonds issued in Europe. One is from Bloomberg, and includes primarily publicly traded firms, while the other is from 9fin, which includes primarily private firms. Theoretically, the dataset from 9fin best represents the universe of European high yield, as the vast majority of new issues entering the markets are from privately owned companies (AFME, 2020). Indeed, private equity firms often attempt to increase leverage on target firms as much as possible, in order to increase return on the equity investment. This strategy, when promoted to the extent that credit rating falls below BBB, will often be punished in the market due to the increasing risk of the default and ensuing risk of a complete equity wipeout, making it unattractive for publicly traded firms. Publicly traded firms also have the option of raising capital on equity markets by the issuance of new shares. Consequently, a large part of high yield bonds from public firms trading in the market were originally issued as investment grade bonds but have since been downgraded due to negative developments in the underlying firm, so-called fallen angels. However, data collection is more easily accessible for publicly traded firms, and more granular data is available. An example of this is accounting data, where all publicly traded firms are obligated to extensive reporting on a quarterly basis. For private firms, this is not always the case, meaning some companies included in the private dataset may have accounting variables from a 2018 annual report as the latest available data. This, combined with an academic interest in studying differences in the effect of text analysis on private and public firms, has led to the decision of using two separate datasets. While the data collection process was largely manual, all data processing and subsequent analysis has been performed using the Python programming language. Throughout the paper, references will be made to the relevant documentation and libraries used within Python, and the programs written will be available as appendix. The following is a brief description of the data collection process for the two datasets.

6.1.1 Public dataset from Bloomberg

The first dataset is collected from Bloomberg, available using company access at Capital Four. Bloomberg stores financial and accounting data on all public companies, and also has a document search function that allows for retrieval of the relevant prospectus for a given bond ISIN. The basis of the dataset was created by listing all active corporate bonds issued in Europe which are trading publicly as of January 1st 2020 (34362 bonds) and applying a filter including only bonds rated as high yield. This yields 3691 bonds. Filters are then applied, in accordance with the literature and the objective of the study. Firstly, banks and other financial institutions are removed, as both the accounting variables and the characteristics of the bonds are completely different, taking the dataset down to 1662 bonds. Then, a filter is applied to only include bonds for which Bloomberg has accounting data for the underling companies, as these are needed to access the credit risk. This takes the dataset to 571 bonds.

Attempting to match a bond with its respective prospectus is a challenging and highly manual process, as the document search feature on Bloomberg is primarily intended to browse through statements from a single company, in order to quickly access company filings, transcripts of quarterly presentations by management, and relevant research by third party analyst firms. As such, searching for a specific document across multiple firms is rather challenging, and not very well-developed. Particularly, ensuring that bonds from a firm which has issued multiple bonds are classified correctly, meaning that each of the bonds are matched to the correct prospectus, is infeasible using the Bloomberg document search feature. From a theoretical perspective, as we are interested in the effect of text in prospectuses on bond performance, the idea of having several bonds from the same firm, which in the majority of cases are tied to the same prospectus (firms often issue multiple bonds in a single offering, or raise additional capital on the same terms as the original bond, also known as tapping the bond), as well as the same accounting variables, also poses a significant risk of simply increasing the weight of firm specific noise in the regression, as two bonds from the same company tied to the same prospectus (often trading at very similar spreads) will have double weight on our results. Therefore, in line with (Kovner & Wei, 2014), we only include a single bond from each company, which is selected based on issue day. This is under the assumption that text features of the prospectus are more relevant for newer prospectuses, as new information from the specific firm and the overall market gradually makes the original prospectus more outdated. We later test this hypothesis by predicting spreads both on issue date as well as on 31-12-2019. The downside of this approach is that companies with latest bond issue after January 1st 2020 are excluded from the dataset. This leaves 191 bond issues from unique firms, all of which are publicly listed. Out of these, 110 had prospectuses available in the Bloomberg document search function. Using Bloomberg's application program interface (API) for Microsoft Excel, we then collect accounting data for the 110 firms, both at the time of the bond issue and as of 31-12-2019. Lastly, in a similar fashion, historical price and spread development for all of the bonds are collected, starting from January 1st 2014. No data for previous years were

available, and since 5 of the 110 bonds are issued prior to January 1st 2014, these 5 bonds cannot be included when predicting spreads at the time of issue, but can still be included when predicting spreads at 31-12-2019.

6.1.2 Private dataset from 9Fin

In addition to the Bloomberg dataset, we were able to collect data on bond issues by firms that are privately held, using a high yield market data provider named 9fin. 9fin describes itself as *"a comprehensive analytics platform that helps fixed income professionals save time and make smarter decisions"* (9fin, 2020), and delivers real time data to firms involved in the high yield industry. With resources from Capital Four, data was collected and processed in similar fashion to the public dataset, yielding a total of 902 prospectuses of which 460 could be matched to an ISIN. Of these 460 prospectuses, manual review showed that 43 of the ISINS matches were duplicates, meaning that two or more issues had been done on a single prospectus, in which case the issue with the largest amount outstanding was chosen. This was done under the assumption that larger issued more accurately predict credit risk as they are more liquid. This could be argued to introduce an issue size bias in the dataset, but as companies usually issue bonds in amounts close to one another, the bias is most likely negligible. Of the 417 remaining prospectuses, 9fin provided accounting data on 369, which constitutes our initial private dataset. It is important to note how the dataset contains multiple issues from the same company, unlike the public dataset. However, in the case of the private dataset, we are able to ensure that all of the bonds are from a unique prospectus, which is crucial for drawing inference from text analysis. As such, we decide to include multiple bonds from one firm. This was a result of a large part of the dataset not being granular enough for actual modelling, further described below. For this reason, a decision of keeping the base dataset as large as possible was made. Additionally, with 293 unique firms represented in the base dataset, the overall universe of private high yield companies in Europe is very well-covered, with no single firm being overrepresented. Another important note about the private data set, is that it covers the entire space of 9fin's data universe, which includes a small amount of public companies. Further details on these splits and more will be provided below. As with the public dataset, historical G-spreads for the bonds were collected from Bloomberg, starting from January 1st 2014. Unlike the public dataset, it was only possible to collect data on the private companies as of 31-12-2019, and for this reason it will only be possible to evaluate the effect on spreads at this date.

6.2 Descriptive data on bond issues

The following page consist of a figure illustrating descriptive data on the two datasets:

Figure 11: Breakdown of empirical dataset by bond issuer & bond characteristics

Apart from the split between private and public, the public dataset of 110 observations and the private dataset of 369 observations appears similar and comparable on all parameters. The following is an in depth description and discussion of all the parameters in question.

Bloomberg classifies the industry of a company through its Bloomberg Industry Classification System, which have been applied for the industry classification of the securities in question. Bloomberg Industry Classification classifies companies on three different levels of granularity; sector, group and subgroup. In order to have enough datapoints for each category, the companies have been categories based on their sector label, which is the highest level of classification available. This decision is also supported by the fact that the number of group and subgroups varies widely between sectors. Energy is divided between oila and gas & coal companies, while consumer discretionary consist of 9 separate groups, ranging from distributors to travel, lodging and dining. All sectors within the Bloomberg Industry Classification System are represented, and no single sector has a large overrepresentation in the dataset.

Firms that operate and headquarter in Europe will be more likely to issue debt on European financial markets, and the country overview is a testimony to that. The country listed is the country in which the underlying company of the security is originated, which explains why countries outside Europe can be represented. However, for both the public and private dataset, it is the case that European countries, particularly United Kingdom, Germany, France, Italy and Spain, as well as the United States, are the main components of the dataset. One potential bias introduced by including all securities issued in a given area, is that securities from companies not native in the given area will tend to concern larger multinational firms, as companies that operate only in their local market are unlike to seek financing on foreign financial markets. An example of this is the representation of a country like Brazil, which is represented through bonds issued by BRF S.A., a Brazilian food company and one of the largest of its kind, and Petrobras, the semi-public brazilian multinational oil corporation. In other words, while the dataset contains mediumsized companies from European contries, it is unlikely that this balance is maintained across countries.

Rating agencies such as Standard & Poors, Moodys and Fitch will usually rate a security before it is issed, although certain private issues are not neccesarily rated. In order to compare ratings across different rating agencies, which employ different scores, we adopt Bloombergs Composite Rating System. The system assigns a rating to a security based on a blend between ratings from DBRS, Fitch, Moody's and Standard & Poor's. The scale ranges from AAA as the highest rating, to D as the lowest. Bonds at or below BBB are categorized as high yield bonds, and bonds above as investment grade. As one would expect, both dataset consists primarily of securities rated BBB- to B- with 4 unranked securities in the private dataset.

The bonds are issued between 2010 and 2019, with 2017 being the most common observation. The public dataset is skewed slightly more towards newer issuances that the private dataset, and

contains more issues dated earlier than 2014. As discussed earlier, a potential explanation for the large number of 2019 issues in the public dataset is the use of latest issue for every firm, meaning that a company which issued debt in 2017 and again in 2019 would only be included as a 2019 entry. However, a certain percentage of bonds will be called or will default in any given period, meaning they are no longer tradeable in the market, and for this reason one would expect, given a fixed amount of issues per year, that the number of bonds currently trading would be larger for years closer to the present. Bond issuance is highly cyclical, however, (AFME, 2020) and for therefore the difference in issuance year observed in the dataset is not in itself alarming.

For both datasets, the majority of bonds are senior secured, and 70% of the public bonds have an inbuilt call clause, allowing the company to redeem the notes at a date earlier than expiry, in return for a premium. 9fin does not provide information on the such call clauses on the notes issued by private companies. Both datasets are composed by a majority of euro and USD denominated bonds, but the while the private dataset only has 5% of the securities denominated in other currencies (primarily CHF), the public dataset has bonds issued in 13 different currencies, which could be an indication that firms large enough to issue bonds in smaller, local currencies are more often than not publicly traded companies.

For the descriptive variables to be used in regression, it is needed to transform them into numerical inputs suited for use as model input. This is done by establishing binary variables, often referred to as dummy variables, for categorical inputs of interest. We first create dummy variables for industry classification, meaning that each bond is assigned variables equal to the total number of industries in the dataset, where all variables take the value 0, except for the variable of the industry which the bonds belongs to, which takes the value 1.

We similarly create geographical dummy variables to control for spread differences across countries. However, as the datasets contain bonds from a wide range of countries, and very few bonds from certain countries, bonds are grouped in four geographical regions: Northern Europe, Southern Europe, United States and Other. This way, effects on bonds issued by European firms can be isolated, and we are able to test for the so-called Southern European premium in accordance with Lie & Nielsen (2015).

To control for differences between bonds issued in different currencies, we also create dummy variables based on currency, with groupings similar to those shown in the figure above: Euro, GBP, USD and Other. We likewise create a single dummy variable controlled for whether bonds are callable or not, as well as a dummy variable controlling for the seniority of the bond. For all variables apart from industry, the first dummy variable is dropped to increase the number of degrees of freedom obtained in the models assessed. This only affects the intuition of the estimator in question and should not alter conclusion. For instance, in the case of geography, only three dummy variables enter the model, and Norther Europe is viewed as the point of origin for any estimators, in the sense that a significant premium (positive or negative) for any other

geography needs to be interpreted as the *difference* between a bond being issued in Northern Europe, and in the geography in question.

For rating, we apply a numerical scale in accordance (Gaillard, 2009). This is done to better capture the fact that ratings are a scale that decreases from better to worse, and that a BB+ bond therefore is better comparable to a BB bond than to, for instance, a CCC+ bond. Non-rated bonds are set as 0, D-rated bonds as 1 and so forth, up to BBB which is set to 14.

6.3 Accounting data

Below is the accounting data (€ M) for the public dataset, with newest data as of 31-12-2019 being displayed. This means that for firms with quarterly reporting, the four most recent reports are combined to a statement for the last twelve months.

Table 1: Descriptive statistic of accounting data on EU Public on an LTM basis as of 31/12/2019

	EBITDA	Adj_Ebitda	Sales	Net_Income	FCF	Total_Assets	Tangible_Asset	total_cash	Net_Debt
Count	110.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0
Mean	1178.31	1292.77	6342.93	365.79	375.38	11379.17	8352.68	974.36	2990.93
Std	2978.7	3087.38	13117.38	925.2	1425.45	24561.24	18286.71	2472.4	7417.49
min	-356.6	-113.45	0.0	-649.11	-1375.65	15.93	2.38	0.31	-7263.0
25%	8.67	13.95	123.57	1.8	0.58	208.26	167.67	11.54	33.66
50%	123.69	144.07	849.96	37.1	16.06	1920.03	1385.49	145.05	381.48
75%	962.92	969.99	5527.79	229.46	304.68	8023.83	6517.42	867.91	2364.78
max	23738.03	22667.53	74731.0	5108.1	12491.12	157398.72	144057.55	19232.0	53990.01

	Total_Debt	LT_Debt	ICR	WC	EV	Market_Cap	Sales_Growth	Sales_CAGR
Count	110.0	110.0	110.0	110.0	109.0	109.0	110.0	110.0
Mean	3964.71	3356.52	2.78	253.39	7862.23	10857.65	0.04	0.08
Std	8735.24	7690.32	5.53	1657.47	16929.06	61944.24	0.15	0.23
min	5.68	1.39	-0.33	-9310.91	0.0	0.0	-0.72	-0.48
25%	72.23	56.07	0.06	-2.15	92.82	42.69	-0.01	-0.0
50%	612.25	478.07	0.34	24.01	954.1	432.52	0.02	0.03
75%	3499.32	2975.05	3.31	319.21	7113.43	5183.28	0.08	0.11
max	59688.59	52695.99	39.32	9344.0	86193.9	642061.73	0.58	1.71

Followed by the same dataset, with most recent financial performance at the time of bond issuance (€ M):

Table 2: Descriptive statistics of accounting data on EU Public on an LTM basis at the date of issue

	EBITDA	Adj_Ebitda	Sales	Net_Income	FCF	Total_Assets	Tangible_Asset	total_cash	Net_Debt
Count	110.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0
Mean	1024.48	1104.39	5979.7	310.59	322.0	9520.5	7486.49	874.61	2148.62
Std	2816.58	2890.73	13294.21	845.71	1410.64	21796.77	18502.33	2337.94	6159.35
min	-468.63	-5.94	0.0	-497.75	-1607.7	0.0	0.0	0.0	-7216.0
25%	7.72	10.03	106.74	2.03	0.07	115.5	110.02	7.34	23.54
50%	86.54	114.01	686.45	23.64	8.44	1782.88	1029.88	73.29	339.78
75%	603.04	630.4	4879.76	199.55	147.62	7343.26	6196.65	617.31	1927.87
max	22535.8	22213.52	73772.0	5193.32	11981.37	157135.94	155452.0	17758.0	53350.73

	Total_Debt	LT_Debt	ICR	WC	EV	Market_Cap	Sales_Growth	Sales_CAGR
Count	110.0	110.0	109.0	110.0	109.0	109.0	104.0	101.0
Mean	3021.04	2441.53	2.47	206.76	7149.66	11644.97	0.04	0.04
Std	7523.62	6305.22	5.46	1941.41	17584.38	68908.27	0.15	0.09
min	0.0	0.0	-0.08	-8110.0	0.0	0.0	-0.4	-0.23
25%	38.73	31.54	0.04	-2.58	98.98	75.77	-0.02	-0.02
50%	537.75	434.81	0.14	19.18	961.82	481.74	0.04	0.03
75%	2715.6	2053.16	2.25	350.74	6121.19	4523.38	0.1	0.09
max	63635.41	54310.41	39.07	13892.0	122126.41	712909.91	0.52	0.3

On average, firms in the public dataset have adjusted EBITDA of 1,1 €bn and a market capitalization, defined as the value of all stock outstanding, of 11,6 €bn. Net Debt, defined as total debt minus cash and cash equivalents, is 2,1€ €bn on range, ranging from -7,2€bn to 53,4€bn. ICR (Interest coverage ratio), defined as interest expense divided by adjusted EBITDA, was on average 2,47.

Notice how for all datapoints, standard deviation is multiple times larger than the mean. This indicates that observations are widely spread, which is confirmed by the large spread between minimum and maximum values. It is also worth noting that, due to the nature of the data, almost all distributions are right skewed as firms usually do not have accounting datapoints far below zero, meaning that a few, large firms will pull up averages of almost all accounting variables. For this reason, the median is often are more indicative description of the overall dataset. For instance, in the case of adjusted EBITDA, the average of 1,1 €bn corresponds only to a median of 114 €m, indicating a small number of larger firms are present in the data (in this case, the Russian energy firm Gazprom with EBITDA well over 20€bn).

It is also clear that several datapoints are outliers. More specifically, it appears that the data contains several null values in places where the true value cannot be 0. As such, missing or

incorrectly filed datapoints are removed for cases where sales, total assets, total cash, total debt and EV or market cap equals 0. Datapoints where net debt is negative is manually confirmed to not be a case of a missing value for total debt plus a positive value for total cash, in which case the data point is removed. Likewise, the maximum value of market cap is several times larger than the maximum value of enterprise value, defined as market cap plus net debt, indicating that a single datapoint is corrupt. Such datapoints are checked manually and removed.

The process of removing outliers and missing data values means that the dataset is reduced, as a entry in the dataset needs values for all parameters in order to be used as model input. As such, two different bonds, where one is missing ICR and another is missing a value for net debt, means that both are removed. This takes the total dataset to 91 datapoints for newest data as of 31-12-2019, and 79 datapoints for data at the time of bonds issuance.

Below is the accounting data for the private dataset from 9fin (€M), with latest data as of 31-12-2019:

Table 3: Descriptive statistics of accounting data on EU Private on an LTM basis as of 31/12/2019

	Sales	OP_margin	Interest_Expens	net_income	NI_margin	adj_ebitda	adj_ebitda_mar	EBITDA
Count	310.0	270.0	286.0	294.0	289.0	249.0	241.0	175.0
Mean	5852.25	0.1	-215.92	-5.12	0.0	1001.66	0.24	740.15
Std	9025.98	0.14	335.78	688.91	0.11	1549.1	0.22	1199.75
min	80.06	-0.4	-2163.4	-3312.0	-0.44	-18.0	-0.01	-18.0
25%	1080.4	0.02	-211.98	-61.66	-0.03	145.94	0.09	123.2
50%	3083.3	0.07	-89.6	10.48	0.01	388.0	0.17	321.0
75%	6668.0	0.12	-47.38	178.59	0.04	1436.1	0.34	816.3
max	94853.3	0.96	24.73	2493.0	0.54	11444.0	1.24	7489.0

	ebitda_margin	Capex	operating_cf	total_cash	net_debt	net_leverage	ICR
Count	167.0	251.0	281.0	312.0	279.0	241.0	229.0
Mean	0.2	-345.57	705.1	492.63	3745.09	2.78	5.17
Std	0.21	505.59	1104.51	1104.02	7684.74	12.03	4.94
min	-0.01	-3061.0	-444.24	-1798.0	-421.4	-100.33	-0.1
25%	0.07	-477.0	84.32	69.96	571.1	2.58	2.26
50%	0.13	-117.17	293.9	139.44	1524.18	3.76	3.46
75%	0.22	-39.41	763.0	557.16	3558.0	5.24	5.96
max	1.24	1.41	6392.0	14179.85	75352.25	20.69	22.36

For the private dataset, companies have an average EBITDA of 1€bn, which is very similar to that of the public dataset. Net debt is on average 3,7 €bn, ranging from -420€m to 75€bn and ICR is on average 5,17. All of these figures are in the same order of magnitude as for the public dataset.

Another similarity is the right skew of all variables, with means being several times larger than medians in most cases, as few large companies skew the distributions.

The data in the private dataset is more consistent, with fewer apparent outliers. Still, the same measures as for the public dataset is applied. It is, however, clear that several companies lack data points for several variables. For instance, only 249 companies have data on adjusted EBITDA, which is crucial as a model input. In order to preserver datapoints, the following measures are applied: In cases where 9fin reports both a ratio as well as the two individual data points that constitute the ratio, the ratio is calculated manually if the two individual data points have more data than the ratio. An example of this is EBITDA margin. With 167 data points, 9fin only reports the EBITDA margin of less than half of the original dataset. To circumvent this, EBITDA margin is calculated manually by dividing EBITDA with sales, which has 175 and 310 datapoints, respectively. Calculated values are then matched against the original ratio to ensure consistency. In the event of EBITDA margin, a decision was made to drop the measure and rely solely on adjusted EBITDA margin, as this measure had 249 datapoints, achieved by dividing adjusted EBITDA with sales. As such, the dataset is reduced from 389 to 176 data points.

Lastly, model inputs for the analysis are calculated. The following ratios are calculated for both datasets:

$$Adjusted\ net\ leverage = \frac{Net\ debt}{Adjusted\ EBITDA}$$

$$FCF\ to\ Net\ debt = \frac{Free\ Cash\ Flow}{Total\ debt\ - Cash}$$

$$Cash\ ratio = \frac{Cash}{Total\ Assets}$$

$$Adjusted\ EBITDA\ margin = \frac{Adjusted\ EBITDA}{Sales}$$

$$Net\ income\ margin = \frac{Net\ Income}{Sales}$$

Here it should be noted that due to data availability, ratios for the private dataset are calculated with two adjustments: FCF to Net debt is calculated using operating cash flow, and cash ratio is calculated as cash divided by sales.

Additionally, for the public dataset, the following ratios are calculated:

$$Equity\ cushion = \frac{Enterprise\ value - Net\ debt}{Enterprise\ value}$$

$$RoA = \frac{Adjusted\ EBITDA}{Total\ Assets}$$

$$WC\ to\ assets \frac{Working\ Capital}{Total\ Assets}$$

$$Asset\ Tangibility = \frac{Tangible\ Assets}{Total\ Assets}$$

$$Sales\ CAGR\ = (\frac{Sales_t}{Sales_{t-3}})^{\frac{1}{3}}$$

Ideally, the private dataset would include a 'Share of Debt' measure as a proxy for equity cushion. However, data for total assets were unavailable for the private dataset. Sales growth were also unavailable.

As described above, in all cases where a preexisting value for the ratio already exists, the preexisting value or the manually calculated value is chosen based on what yields the largest amount of data points. In these cases, manually calculated and preexisting values are matched in order to ensure consistency. The ratios are the final input that enters the models in, as they in theory are equally scaled between companies, which is not the case for non-ratio accounting variables such as sales and adjusted EBITDA.

A special consideration needs to be made in the few cases of negative EBITDA. While there is nothing theoretically wrong with a company being negatively levered in the case where cash outweighs total debt, having a company with a net leverage ratio far below zero, simply because EBITDA is slightly below zero, is not meaningful. Generally, it is nonsensical to refer to leverage ratios of a company with negative EBITDA, and as such, these values are removed.

Below is the descriptive statistics for the model inputs are presented for the private dataset:

Table 4: Descriptive statistics of accounting ratio variables (EU Private)

	Net leverage	FCF / Net Debt	Cash ratio	Adj. EBITDA Margin	NI Margin
Count	176.0	176.0	176.0	176.0	176.0
Mean	4.13	0.3	27.88	0.23	0.0
Std	1.85	0.79	73.49	0.21	0.12
min	0.05	-0.27	1.18	0.02	-0.44
25%	2.94	0.11	7.43	0.09	-0.04
50%	4.06	0.19	12.5	0.15	-0.0
75%	5.25	0.29	25.48	0.31	0.04
max	10.22	10.12	651.37	1.24	0.54

And for the public dataset at 31-12-2019:

Table 5: Descriptive statistics of accounting ratio variables (EU Public)

	Adj. EBITDA Margin	Adj. EBITDA Margin(i)	Asset Tangibility	Asset Tangibility(i)	Cash ratio	Cash ratio(i)
Count	91.0	79.0	91.0	79.0	91.0	79.0
Mean	0.23	0.21	0.77	0.79	12.72	17.95
Std	0.17	0.16	0.2	0.2	16.43	37.91
min	0.02	0.01	0.15	0.15	1.28	0.86
25%	0.11	0.1	0.67	0.66	4.29	4.33
50%	0.18	0.16	0.81	0.84	7.65	8.95
75%	0.3	0.26	0.94	0.93	13.22	14.32
max	0.82	0.81	1.0	1.0	93.23	291.75

	Net leverage	Net leverage(i)	RoA	RoA(i)	WC / Assets	WC / Assets(i)
Count	91.0	79.0	91.0	79.0	91.0	79.0
Mean	2.6	3.98	0.13	0.12	0.06	0.06
Std	4.0	12.71	0.06	0.06	0.13	0.13
min	-15.55	-1.07	0.01	0.0	-0.36	-0.25
25%	1.38	1.36	0.1	0.08	-0.01	-0.02
50%	2.55	2.34	0.11	0.1	0.05	0.06
75%	3.31	3.6	0.14	0.13	0.15	0.13
max	33.19	114.5	0.42	0.36	0.32	0.34

	Equity cushion	Equity cushion(i)	FCF / Net Debt	FCF / Net Debt(i)	NI Margin	NI Margin(i)
Count	91.0	79.0	91.0	79.0	91.0	79.0
Mean	0.61	0.68	0.55	0.14	0.07	0.06
Std	0.36	0.2	4.09	0.3	0.09	0.08
min	0.04	0.22	-1.54	-0.82	-0.08	-0.22
25%	0.39	0.57	0.03	0.01	0.02	0.02
50%	0.58	0.66	0.13	0.12	0.05	0.05
75%	0.73	0.81	0.26	0.27	0.09	0.1
max	2.69	1.46	38.99	1.36	0.48	0.36

In line with expectations, we find that firms in the private dataset are generally higher levered than firms in the public dataset (4,13 vs. 2,60) which would be the case if a larger portion of the private dataset contains issues in relation to PE or LBO deals. Cash generation and profitability are almost identical

6.4 Text analysis of bond prospectuses

The bond prospectuses carry huge amount of information. Many of them contain hundreds of pages of information on everything from important comments to the financial numbers reported, management's plan for the future of the company, the use of the proceeds from the bond offering, and the risk that could potentially affect the company in the future. While a lot of this information is perfectly readable for humans, often investment analysts browse through the prospectuses for important information as part of their due diligence process, it is very difficult for a machine to extract any information from textual data. Text data is unstructured data, and does not fit neatly

into rows and columns, which is the typical data structure adopted in machine learning models and general quantitative analysis. To go from a large body of unstructured text to useful input for machine learning, the texts need to be processed using several steps of natural language processing. In short, natural language processing (NLP) is a cross-field between linguistics and machine learning, concerned with the interactions between how humans use and understand text and how computers can understand text (Bird et al, 2009). The following section will outline the natural language processes deployed in this paper to turn the corpus of prospectuses into the text variables used in the models

6.4.1 Text pre-processing

6.4.1.1Parsing the text

The first step of the process is to extract the raw text data from the corpus of prospectus PDF files. Most coding languages, python included, cannot simply crawl through and process information stored in PDF formats. To access the text, a PDF to text parser had to be deployed on the data. We use the python PDF parser library Tika (Version 1.6), to crawl through the documents and parse them into raw text, which was then stored in txt files. The PDF to text parsing works by decoding all of the PDF text and turning it into Unicode and then turning it back into readable text using UTF-8 Encoding. For the private dataset, we lost 4 datapoints as only 898 of 902 prospectuses could be successfully parsed. All 110 prospectuses for the public dataset was parsed successfully.

As described in the theoretical considerations behind model input, informative content of prospectuses is often found in specific sections. In particular, the risk section, MD&A and Business & industry description. For this reason, ideally the analysis is performed on one or more of these sections, while more generic sections, which often include great amounts of legal boilerplate language, are removed. Prospectuses collected from Bloomberg contains clear section encoding, which allows for stripping of individual sections. This revealed that the only section which all prospectuses have in common is a risk section[9]. We therefore use risk sections as unit of analysis for the public dataset, while the entire prospectus is processed for the private dataset as no section encoding was available.

6.4.1.2 Tokenization

With the raw text extracted, the next step was tokenization of the data. Tokenization is the process of cutting the raw text strings into individual words and sentences, called tokens, which can be used as unit of analysis. We used the NLTK (Natural Language Tool Kit), which is one of the most widely adopted NLP python libraries for the tokenization of the texts (Bird et al, 2009). We parsed the text for both word and sentence tokens but decided in favor of using word tokens as the unit of analysis for the further process of the textual data.

[9] Management discussion were only present in 67% of prospectuses, and industry & business description were only present in 31%

6.4.1.3 Token pre-processing

With individual words tokenized, the tokens must be further processed to ensure conformity. We employ four widely used token processing measures: lower-case conversion, remove punctuation, stripping-white space, and stop-word removal (Bird et al, 2009). The Python language is case sensitive, as it will see the same word but with different letters capitalized as different words. While both case sensitivity, punctuations and whitespace, such as newlines or spaces, can contain important informational distinctions when dealing with text on a sentence level, when creating word tokens, you want every word to be categorized identically regardless of case. For instance, both 'Bond', 'bond', and 'bond.' should be tokenized as the same token. Whitespace or case sensitivity therefore does not account for any useful distinction and would simply add noise in the corpus of tokens. Similarly, stop-word removal is applied to get rid of tokens without any informational value. Most tokens of any text in the English language will be a group of articles, pronouns such as 'and, or, too, the' etc. These tokens will be highly prevalent in all texts while not holding any information on a tokenized level. While there is no universal list of stop-words, as what you consider a stop-word can vary from field to field, we apply the wordlists from the two python libraries NLTK and Sci-Kit Learn (SKLearn), which are both widely adopted in the NLP field (Bird et al, 2009).

The two last steps applied in the processing of tokens are lemmatization and n-gram tagging. Lemmatization is the process of reducing a word to its base form, more specifically the dictionary form known as lemma (Bird et al, 2009). All verbs in different tenses are turned into present tense, nouns are changed into singular, synonyms are unified etc. Lemmatization takes into account the context of the word, which allows it to discriminate between identical words with different meaning depending on the context. i.e. 'post' is a verb in the sentence "I post the mail" and a noun in the sentence "a post about the covid-19 crisis on reddit". For proper lemmatization we applied the NLTK python library stemming algorithm (Bird et al, 2009). For the most computational advanced text feature, the LDA topic modelling, we also applied N-gram tagging for the tokens. N-gram tagging is a contiguous sequence of n words from a text. When tokens are just single words without context, we have unigram tagging. To achieve higher granularity on the topic in the topic modelling we applied bigram tagging, where all words are tagged pairwise, and used as tokens in context of one another.

6.4.2 Vectorization – turning pre-processed text into machine learning language

After pre-processing the data, the second step applied to turn the text into input for machine learning is to vectorize the data. When vectorizing the data, we turn every text into one vector of integers representing the tokens in the text. This is known as a Bag-of-Word (BoW) representation of the data, which will be a matrix consisting of rows according to every document in the entire corpus and columns according to each token found in the full corpus of texts (Daumé, 2017). For each time a token appears in the document it will get an integer value in its vector according to the respective token, so if the token 'bond' appears 10 times in the document its

vector will hold the integer 10 in the 'bond' token column. A common downside to BoW data representations is that if the texts share some characteristics, some word will be very common across all documents, getting a high BoW score, while not containing much predictive value. Similarly, stop-word-like words not removed will also provide noise in the dataset, as they will be very common. These flaws of the BoW data structure were also present in both of our prospectus datasets, with words such as 'notes', 'issuer', and several numerical figures scoring high BoW scores.

6.4.2.1 Term Weighting: TF-IDF

This issue can be resolved by applying a token weighting to the dataset (Emadzadeh et al, 2010). We apply the TF-IDF weighting scheme, which is widely applied as weighting scheme for text prediction in the literature (Chen & Schumaer, 2007; Loughran & McDonald, 2011; Fung et al, 2009; Alberg et al, 2007). TF-IDF stands for *Term Frequency – Inverse Document Frequency*, and is a technique to vectorize words in a document of a corpus, where the weight of each word corresponds to its importance in both the full document and in the corpus as a whole (Daumé, 2017). To capture the weight of both the importance of a token in the document and the importance of the token in the whole corpus both the *Term Frequency* and the *Inverse document frequency* is calculated (Daumé, 2017):

$$TF - IDF = Term\ Frequency * Inverse\ Document\ Frequency$$

Term Frequency is the frequency by which a token appears in the document. It counts every appearance of a token and store it as an integer as in the BoW model, but in order to not giver higher importance to long documents by default it divides the count of each token by the total amount of words in a document:

$$TF(t,d) = \frac{Count\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

Tokens that appears many times in a document compared to the length of the document will have a high TF score. While this term takes the overall length of documents into account, it still doesn't take into account that some words may be very common and appear many times in almost all documents, and thus don't hold much predictive value (ibid). Therefore, an IDF score is calculated for every token in every document as well. The IDF score measures the importance of the token across the full corpus of documents. The document frequency is the occurrence of the token t in the set of N documents. To normalize this number and make it comparable across data sets the occurrence of the token across documents is divided by the number of documents. This number is then inversed to be: $\frac{N}{df}$ for high values of N and low document frequency, this number will explode, so to normalize it and dampen the range we take the log of the expression (ibid):

$$IDF = Log(\frac{N}{df + 1})$$

Thus, tokens apparent in almost all documents will have a very low IDF score, tokens apparent in all documents will have an IDF Score of very close to 0, while tokens apparent in only a few documents will have a very high IDF score. The full TF-IDF score will then be calculated as the following two expressions multiplied:

$$TF - IDF = \frac{Count\ of\ t\ in\ D}{number\ of\ words\ in\ d} * Log(\frac{Number\ of\ documents}{number\ of\ documents\ tokken\ appears\ in + 1})$$

The TF score will be high for tokens apparent many times in the document, while the IDF score will be low if this token is also apparent in most other documents. High TF-IDF weights will thus be given to tokens that appear many times in a specific document but not in most other documents (Ibid).

6.4.3 Feature generation

With the text pre-processed and vectorized using TF-IDF, the next step is to analyze the text data and turn it into features, or input variables, that can be used in quantitative models. One way of using the text input is simply to use the TF-IDF vectors as raw input and each score as a feature. This is a common approach on huge datasets and resembles the way image recognition algorithms classify pictures based on a Bag-of-pixels (ibid). However, since each text is very long containing many unique tokens, and the overall number of datapoints in our dataset is limited, we would end up with many more features that datapoints, which is not best practice for machine learning (Ibid). Instead we had to analyze the textual data using other techniques more relevant to the study.

We compute three different measures of textual features which can be used as input. The first measure is a fundamental text content score for each prospectus, based on pre-constructed wordlists which are used to score the fundamental composition of certain content in the text. The second measure is sentiment analysis of the sentiment and subjectivity found in the texts. The third measure is a semantic approach based on unsupervised clustering of the tokens using Latent Dirichlet Allocation (LDA) to generate topics across the corpus and a weight towards each topic for each document.

6.4.3.1 Fundamental score based on wordlists

Several scholars have used wordlists to analyze the content or the tone of voice of a text. Early papers, such as Tetlock (2007), Feldmand et al (2008), Alberg et al, (2007), used off-the-shelf dictionaries from other disciplines, such as the Harvard IV dictionary (which is a general dictionary categorizing words. It contains a list of 'negative' and 'positive' words etc.). However, as more thoroughly explained in in the literature review, Loughran & McDonald (2010) found that such off-the-shelf dictionaries tend to wrongly classify words in financial documents. In their 2010

study three fourths of all the words classified as 'negative' by the Harvard dictionary, are typically not considered negative in a financial context. To solve this problem Loughran & McDonald manually constructed a wordlist of words with strong indicative value in financial text, which they use to analyze their corpus of IPO prospectuses. Their wordlist contains words belonging to four different types of indicative words: 'Negative words', 'Positive words', 'Uncertainty words' and 'Litigious words'. Their wordlist has been accepted and applied by other researchers in the financial domain (Bartov, 2011; Deokar & Tao, 2015). For our fundamental score we use the wordlist constructed by Loughran & McDonald. We construct a fundamental score for each of the four categories of words in Loughran & McDonald's wordlist, by first counting the number of 'negative', 'positive', 'uncertain' and 'litigious' words apparent in each document, and control for the length of the documents by dividing by the total number of words in the document.

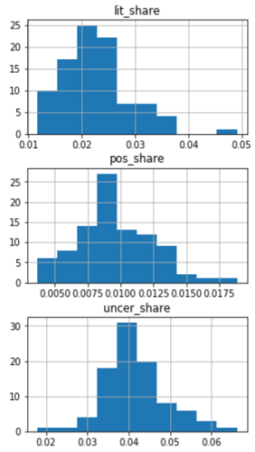
This creates four fundamental scores: 'negative share', 'positive share', 'uncertainty share', and 'litigious share' for each document. Along the lines of previous studies (Tetlock, 2007, Tetlock, 2008, Bartov, 2011, Alberg et al, 2007) We standardize these measures by subtracting each of the individual document fractions by the fraction mean across the corpus of text and dividing by the standard error of the fractions.

$$Negative\ score_i = \frac{Neg\ share_i - \mu}{\sigma}$$

$$Positive\ score_i = \frac{Pos\ share_i - \mu}{\sigma}$$

$$Uncertainty\ \ score_i = \frac{Uncer\ share_i - \mu}{\sigma}$$

$$Litigious\ score_i = \frac{Lit\ share_i - \mu}{\sigma}$$



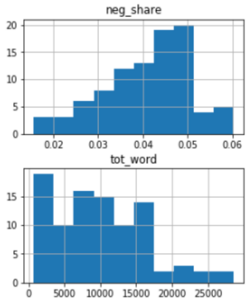


Figure 13: Histogram of wordcount scores for EU Private

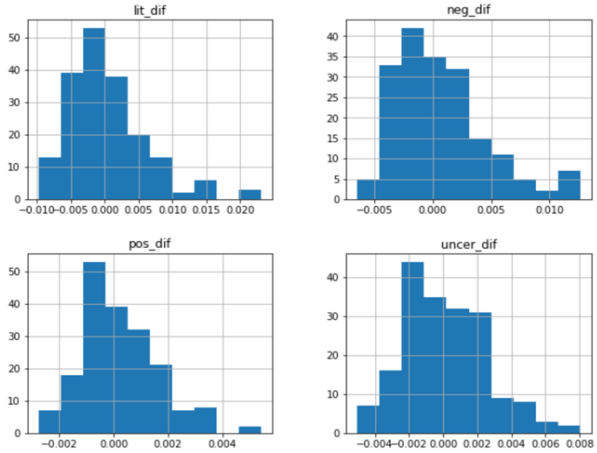


Figure 12: Histogram of wordcount scores for EU Public

These four scores are then used as text features in our quantitative models. We print histograms of the wordcount scores to check the distribution of the data, and to check for any signs of misprocessed or irregular data:

Both the wordcount scores for the public and the private dataset seems to approximately follow a normal distribution centered around the means of the distribution and with no extreme outliers or other spikes in the data distribution. This is very common for a lot of variables and could therefore be expected for a measure such as the wordcounts (Stock & Watson, 2015). Of the wordcount scores only the negative score for the public dataset has a significant left skew in the data, indicating that more text will have slightly less use of negative words compared to the mean than there will be companies with a higher use of negative words than the mean. However, the negative score for the private dataset is skewed to the right, which could indicate that this score also follows a normal distribution and the skews are simply a result of the limited size of the datasets, which can lead to some variance in the dataset distributions compared to the true underlying distribution (Stock & Watson, 2015).

6.4.3.2 Sentiment Analysis

As the second set of features we conducted a sentiment analysis of each of the prospectuses. While simple in theory, sentiment is a powerful tool, which can be applied to huge volumes of unstructured data in order to convert it to a scale or a category that you can more easily interpret, understand, and work with (Farhadloo & Rolland, 2016). Sentiment analysis is a text analysis method used to score the polarity or the sentiment tone in any given text. Different sentiment analysis algorithms output the sentiment as a continuous score (e.g. from -1 to 1) or as different classes (e.g. positive, negative, neutral etc). By conducting sentiment analysis on the documents, we can turn the huge volumes of text into a few features summarizing the sentiment derived from the full body of the texts. These features can in turn be used as model input.

The three main types of sentiment analysis models and algorithms are: rule-based, automatic and hybrid algorithms (Farhadloo & Rolland, 2016). Rule-based algorithms works similarly to the

wordlist method applied above. Two list of lemmas of polarized words (either very positive or very negative) words are created on beforehand, and the algorithm is then looping through the documents, and counts the number of sentences, n-grams, PoS tags etc. which contains words on the positive or negative list and averages a score across the document. The upside to using such a model is that what the algorithm does is very clear and structured, and the score outputted is very interpretable. The downside is that sentiment is a highly subjective. It is estimated that people only agree 60-65% of the time when determining the sentiment of a text (Ibid). An automated sentiment analysis algorithm uses no predefined wordlist or library. Instead it uses machine learning to loop through large volumes of labelled text[10], and then create rules based on the structures discovered in the data. Such algorithms are less subjective and can be more sophisticated but have the risk that the type of text they are trained on is of a very different character than the text they are subsequently applied on (Ibid). Hybrid algorithms combine the desirable elements of the two other types, by both creating some clear pre-determined rules to ensure consistency across different types of text, but also training an algorithm on a corpora of text (Ibid).

For sentiment analysis we chose to apply the algorithm from the python library TextBlob. We chose to apply the TextBlob sentiment analysis for several reasons. Firstly, it is a widely adopted and used sentiment analysis algorithm (Shekhawat, 2019; Banati et al, 2017). Secondly, it is a hybrid algorithm, which has both been programmed with some core rules, but have also been trained on huge corpora of text and is continuously updated (TextBlob Documentation, 2020). Thirdly, TextBlob is built on both advanced Parts-of-Speech tagging and N-gram tagging which enables it to handle both 'negation', i.e. it can distinguish between the positive 'great' and the negative 'not great'. It also handles modifier words, such as 'very' or 'some' etc. which it uses to amplify its scores. The output of sentiment analysis using the TextBlob algorithm is two scores: *Polarity* and *Subjectivity.* Polarity aims to capture the sentiment and is a float value within the range [-1.0 to 1.0] where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment. Subjectivity is a float value within the range [0.0 to 1.0] where 0.0 is very objective and 1.0 is very subjective. Subjective sentence expresses some personal feelings, views, beliefs, opinions, allegations, desires, beliefs, suspicions, and speculations whereas Objective sentences are factual (TextBlob Documentation, 2020). These two scores for each document are used as our second set of text features. While these two features are both potentially more sophisticated than fundamental scores from wordcounts, as they can better understand the context the words appear in, they also hold a higher potential for error or bias. A potential source of error or bias in these two features is the fact that while the TextBlob algorithm has been trained on a wide variety of text, we apply it only on text within the language of financial

[10] A commonly used example is reviews which has stars as the label.

statements, which may vary greatly from the tone and language used in the bulk of the text that the TextBlob sentiment analysis algorithm has been trained on.

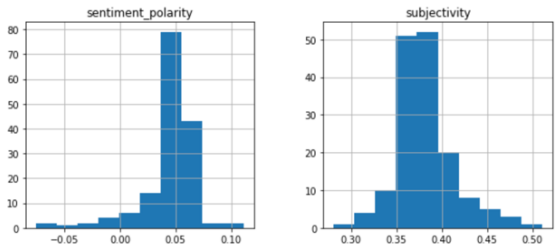


Figure 14: Sentiment Analysis EU private: full prospectuses

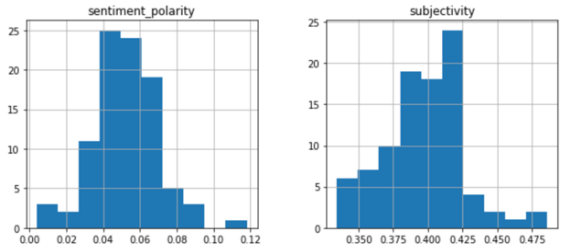


Figure 15: Sentiment Analysis EU public: Risk sections

As with the wordcount scores we check the histograms of the polarity and the subjectivity scores. Figure 14 shows the sentiment and subjectivity score for the EU Private dataset, where the unit of analysis is the full prospectuses and figure 15 shows the histogram of the sentiment and subjectivity score of the EU public dataset, where the unit of analysis are the risk sections only. From figure 14 we can observe that for the full prospectuses on the EU private dataset, both the sentiment and the subjectivity is highly centered around the mean. This could indicate that the language used for most of the prospectuses are written in the same style of language using the same words. As large part of a prospectus are standardized sections of financial commenting legal disclaimers, one can expect these parts to have the same tone of voice across prospectuses, which the data seems to prove. The EU public dataset, seem to be more resembling a normal distribution, which also makes intuitive sense as the content and the use of words for this section is less standardized than the many other parts of a prospectus, such as the financial commenting. One thing to note is that the dispersity in the EU public data is not very high (Polarity ranges from 0 – 0.12, and subjectivity ranges from (0.325 – 0.5). This means that all of the prospectuses are written in language that is very standardized.

6.4.3.3 Topic Modelling using LDA

The last set of text features computed for the corpus of prospectuses is topic model weights using Latent Dirichlet Allocation (LDA). Topic Modelling is an unsupervised machine learning method, that cluster corpora of text around a number of fixed topics, without any supervised classification for what those topic should be, similar to how many clustering algorithms for quantitative data work (Tufts, 2019). Topic modelling holds potential for unlocking information about the text that the other two sets of features are not designed to capture. It can find themes beyond that of positive and negative connotated words and sentences (or uncertain or litigious), including hidden themes or clusters found in the corpora of documents (Blei et al, 2003). To perform topic modelling on the corpus of prospectuses we used LDA, which is one of the most popular topic

modelling methods (Blei et al, 2003). LDA works under a series of assumptions and conditions, which will briefly be outlined below.

LDA works under the assumption that each text is a 'Bag of Words'. This means it does not consider in which order the words appear, or the grammatical role of each word. For the algorithm to work properly it therefore needs to be fed text data that has been pre-processed through the steps mentioned above (tokenized, lower-case converted, removed stop-words, lemmatized) and vectorized in the form of a BoW or a TF-IDF matrix. Also, words that appear in most documents will carry little informative value and will only make the clustering mechanism heavier (Tufts, 2019). Thus, stop-word removal, and TF-IDF weighting makes the algorithm perform better. The topic must be known or guessed beforehand. The algorithm can only cluster the data around a pre-specified number of topics. While there is no exact right number of topics, for the algorithm to perform well, you need to set the number of topics to a level that approximately matches the patterns in the data. If the data has no intuitive number of topics, as is the case in this study, common practice suggest using trial and error to determine a suitable number (Tufts, 2019). LDA works by having each word assigned to both a document, which is information known beforehand, and to a topic. As the topic is not known on beforehand it is introduced as a hidden layer, known as a latent. The weight towards the latent needs to be calculated. When each word is assigned to a document and a topic, the weight of a document towards each topic can be calculated (Tufts, 2019).

The LDA Algorithm starts by assigning each word in each document randomly to one of the k prefixed topics. Then it calculates the proportion of words in document d that are assigned to topic t: $p(topic\ t \mid document\ d)$ and the proportion of assignments to topic t over all documents that come from this word w: $p(word\ w \mid topic\ t)$ (Kulshrestha, 2019). LDA then represents a document as a mixture of topics and a topic as a mixture of words. If a word has a high probability of being contained in a topic, then all documents with a high frequency of that word has a high probability of being in that topic. For each of the proportions above achieved from the random assignments the algorithm then updates the probability of a word belonging to a topic through: $p(word\ w\ with\ topic\ t) = p(topic\ t \mid document\ d) * p(word\ w \mid topic\ t)$ (Kulshrestha, 2019). This process is iterated until the probabilities that a word belongs to a certain topic cannot be enhanced further.

The LDA Topics and the corresponding topic weights for each document are computed using the Gensim python library. Gensim is a commonly used NLP library which is specialized on topic modelling, which is why we chose it over more generalized libraries, such as Sci-Kit Learn (Gensim Documentation, 2020). We ran the topic modelling a few times with different numbers of topics to check what yielded the best results. We ended up with five topics for the EU private dataset, as the unit of analysis for that dataset was the full prospectus, which mean that more different topics can be dominant across the prospectus. For the EU Public dataset, where the unit of analysis is the

risk sections only, we only ran the LDA analysis with three topics, as less topics can be assumed to be present amongst the risk sections only than amongst the full prospectuses. We found these numbers of fixed topics to give the best results, represented in figure 15 and 16

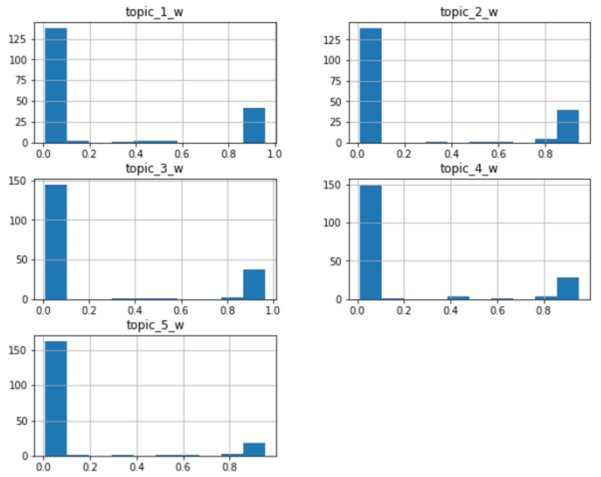


Figure 17: Topic Weight distribution: EU private

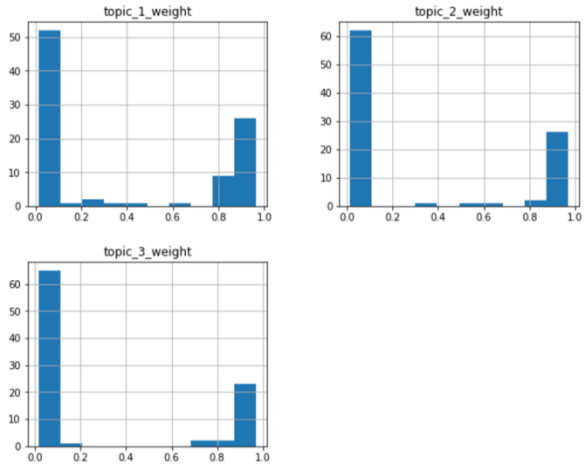


Figure 16: Topic Weight distribution: EU public

Figure 15 shows the Topic Weight distribution for the EU private dataset and figure 16 shows the Topic Weight distribution for the EU public dataset. From the histogram it is evident that the LDA algorithm has successfully been able to associate each document with a distinct topic, as almost all documents have either topic weight close to one, meaning this document belongs to this topic, or a topic weight close to 0, meaning the document does not belong to the topic. If the number of topics had been set to high or two low or the underlying document had been to identical, the algorithm would not have been able to create such distinct topics, which would result in much more uniformly distributed topic weights (Tufts, 2019)

6.4.3.4 Correlation of the text features

After all of the text features have been calculated, we run a correlation analysis, to control for perfect multicollinearity. If some of the text features captures the same underlying pattern of a text, they would yield very high correlation scores and it would be best to leave one of the features out of the models.

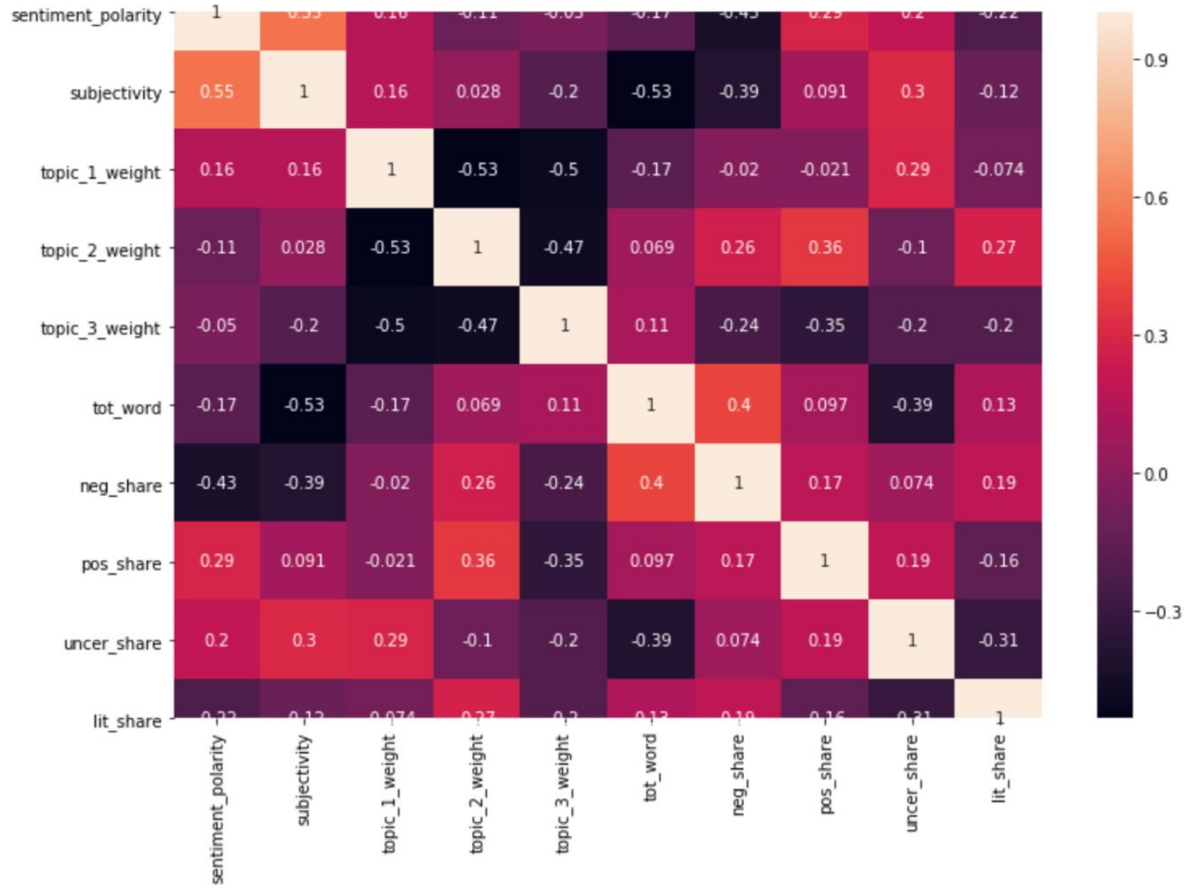


Figure 18 Correlation matrix of text features

Figure 17 shows a correlation matrix of the pairwise correlation amongst the text features computed. Across the different text features no correlation is close to either 1 or -1, so we can rule out the case of perfect multicollinearity. However, a few correlations are worth commenting on. The *negative score* and *Sentiment Polarity* are negatively correlated with a correlation coefficient of -0.43 meaning that a higher share of negative words leads to a lower (or more negative) polarity score. This is in line with what you would expect as there should in theory be an overlap between the two metrics, as a lot of the negative words on Loughran & McDonald's (2011) wordlist, would also trigger the TextBlob sentiment analysis algorithm as negative. A coefficient of -0.53 of the pairwise correlation between the total number of words and the subjectivity scores, indicates that documents that are written more subjective also tends to be shorter.

6.4.4 Text processing - Review

Figure 18 depicts the full text processing flow from the raw data collection to the input of the final predictive model. First the prospectuses are downloaded in PDF or HTML form from the respective database, in this paper we used Bloomberg and 9Fin, but if you want to replicate the study on American high yield bonds, EDGAR could be a good potential database to collect the raw

prospectus data from[11]. Secondly, the PDFs and HTML files are parsed into raw text strings which can be read by the python computing language. The raw text strings are then pre-processed and followingly vectorized and turned into a TF-IDF Matrix. From the processed and vectorized text data the features regarding aspects of the text are computed to function as input for the final predictive models used to predict the risk spreads of high yield bonds. Every step except the data collection is written in python scripts, which means that every step after the data collection could easily be automated from the setup created in this paper. However, the data collection was done in a very manual manner, which proved to be highly time consuming, and left us with a limited sized dataset. If this paper was to be replicated with a much larger dataset, or if a company such as Capital Four Management would adopt this model and use it on an operational level where it would be kept continuously updated, one would have to automate the data collection through the construction of a database crawler, similar to the setup developed by Deokar et al (2018). However, the development of a database crawler for the purpose of this paper was considered to be out of scope.

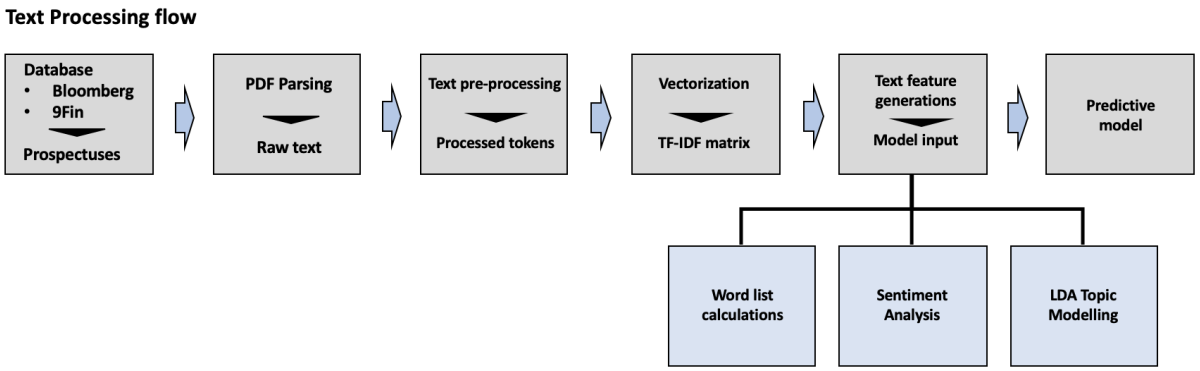


Figure 19: Text processing flow

6.5 Bond performance data

As discussed, several measured for assessing bond performance exists. The three most relevant are yield, G-spread and Z-spread. Below, we present the time development of these measures, as well as the average closing mid-price of the bonds in the dataset:

[11] EDGAR, or *Electronic Data Gathering, Analysis, and Retrieval system* is an SEC database

Public dataset:

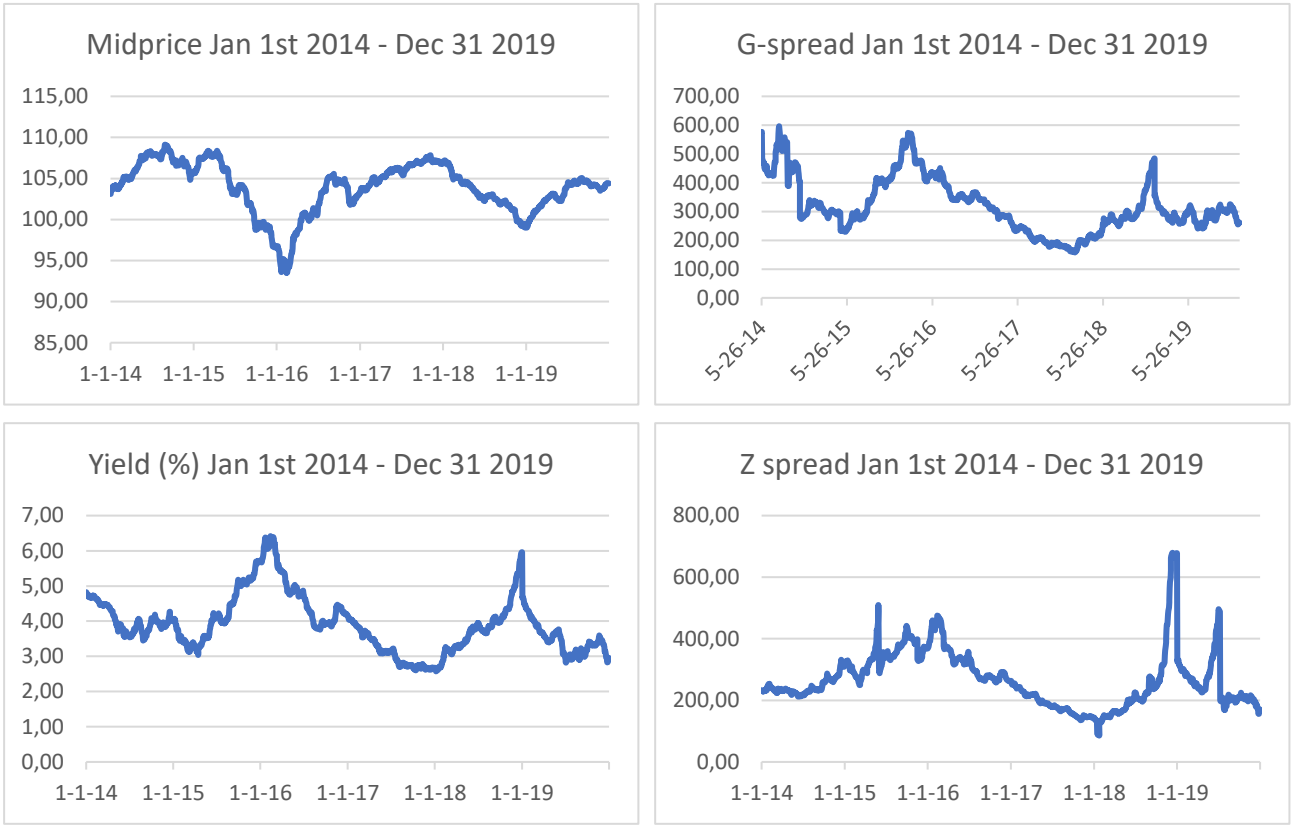


Figure 20: Price, G-spread, Yield-To-Worst, Z-spread of the EU public bond dataset

Private dataset:

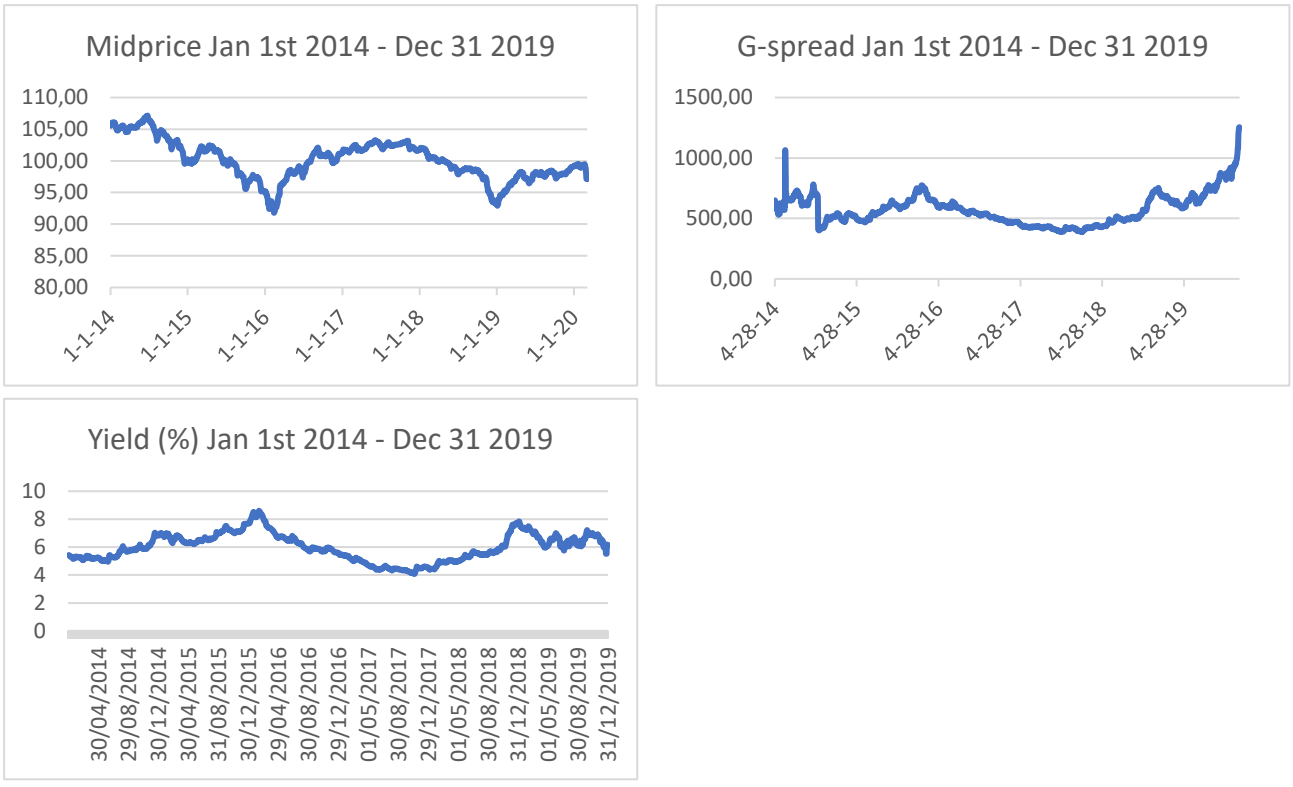


Figure 21: Price, G-spread, Yield-To-Worst, Z-spread of the EU private bond dataset

For both the public and private dataset, bond prices started 2014 by gaining momentum, before experiencing substantial drops in 2015. As markets bounced back in 2016, average prices rose above par, at which level they traded until end 2019. As would be expected, the graph for yield to worst, G-spread and Z-spread show a development inverse to prices. This development is very similar to that of the leading European high yield index, HPC0 (FRED, 2020), indicating that the dataset is a good representation of the overall high yield market. A few interesting observations can be made from the charts. The curve for G-spreads as well as yield to worst appear to be shifted upwards for private bonds, indicating (under the assumption of otherwise similar firms in the two baskets) that a spread premium for private firms is present. Also, towards the end of the analysis period, a sharp drop in price and corresponding increase in G-spread is noted for private firms, which appears not to be matched in the public dataset. A deeper look into this development reveals that deviations are caused by a very small number of bonds with abnormal values, which can cause dramatic swings, particularly in spreads. An elaboration on extreme yields, particularly in the private dataset, will be provided later in this section.

The objective of the study is to study the effect of textual features of bond prospectuses on bond performance. Since textual features of bond prospectuses are static in time, designing the dependent variable in a similar fashion is appropriate. As such, in order to produce suitable model input, values each of the four variables are taken for two separate dates: The date of issue, meaning the closing value of the first trading day, and the closing value of the last trading day of 2019, labelled 31-12-2019.

The choice of cutoff date was made with several considerations: Firstly, a large majority of companies use calendar year reporting structure, meaning that fourth quarter terminates ultimo December. This means that no preliminary earnings would have been released, and consequently incorporated in prices. Had a more arbitrary date been choosing, some firms might have published reporting while others have not. This is mainly an academic concern to keep variables as comparable as possible, as a successful spread predicting model naturally needs to produce consistent results regardless of the day of the year. Secondly, the recent global pandemic of 2019-nCoV acute respiratory disease, has severely shaken financial markets in previously unseen magnitudes, as illustrated by the following graph, containing the same data as figure 20, but also including data from January 1st 2020 to April 20th 2020:

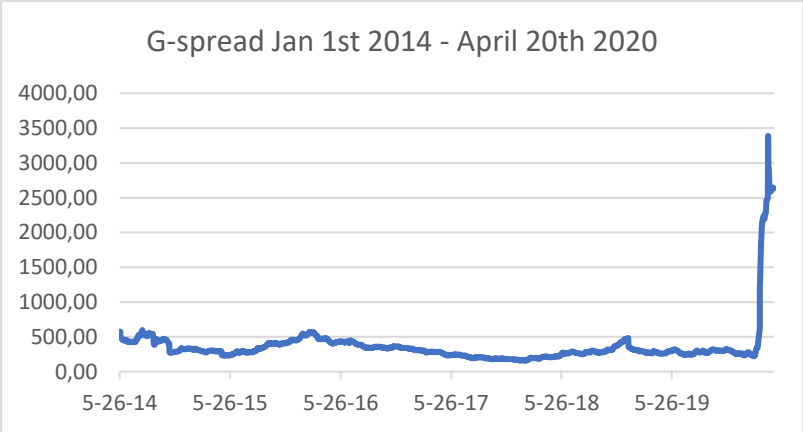


Figure 22: G-spread of EU public bond dataset before and after Covid-19 Crisis

This severity of the volatility depicted above cannot be understated, and it is fair to assume that any regression attempting to predict spreads would simply indirectly predict corona sensitivity. While previous studies on equity markets have employed ways to control for unseen turmoil, particularly the 2008 financial crisis, this is beyond the scope of this study, and a more feasible solution is the use of ultimo 2019 as cutoff date.

Below the descriptive statistics of the values are presented for the private dataset (recall that 9fin only records most recent accounting data as of 31-12-2019, and for this reason, only bond performance data as of that date is presented):

Table 6: Bond performance data for the private dataset

	Issue_size	coupon	Midprice	Yield	Z Spread	G Spread
Count	369.0	369.0	363.0	323.0	200.0	304.0
Mean	429.92	5.07	97.18	462.69	390.77	477.82
Std	271.59	1.96	18.49	667.95	694.85	698.2
min	50.0	1.12	1.48	-333.6	-678.81	-640.47
25%	260.0	3.62	98.57	86.2	89.07	136.37
50%	375.0	4.75	102.55	278.4	213.17	271.05
75%	500.0	6.25	104.77	559.85	461.89	544.52
max	2750.0	11.75	115.04	4709.9	4746.89	5337.02

And for the public dataset:

Table 7: Bond performance data for the public dataset

	Issue_size	Coupon	Midprice	Iss. Midprice	Yield	Iss. Yield	Z Spread	Iss. Z Spread	G Spread	Iss. G Spread
Count	110.0	110.0	110.0	110.0	110.0	110.0	70.0	108.0	105.0	105.0
Mean	596.16	4.25	102.84	100.3	2.55	4.12	145.69	289.96	249.08	335.59
Std	322.7	1.91	9.09	3.21	4.52	1.89	527.19	241.91	396.82	157.48
min	75.0	0.88	55.8	88.59	-21.01	0.0	-2503.16	-57.1	-1507.14	-48.83
25%	400.0	2.8	101.25	99.57	0.92	2.77	81.34	176.65	127.09	230.51
50%	500.0	4.06	103.89	100.48	2.22	3.98	124.5	249.4	186.2	315.62
75%	750.0	5.39	107.03	101.32	3.74	5.16	214.11	348.6	271.02	414.69
max	2000.0	9.25	123.8	113.7	20.9	8.32	2116.43	2207.77	2149.66	855.12

The average issue size of bonds in the dataset is 430 €m for the private dataset, against 596 €m for the public dataset with comparable spread in the distribution. Coupon, meaning the nominal percentage value of the bond coupon, is also comparable across private and public bonds, with private bonds on average paying slightly larger coupons. It is important to note that this measure is not in itself very telling, as the number for some securities represents a spread premium over the regional inter-bank interest rate, whereas the measure for other securities present the actual percentage value of coupon payments. As such, without controlling for fixed versus floating rate notes through dummy variables, including coupon as a model parameter has little theoretical underpinning. Another parameter which is disqualified as a model parameter is Z-spread, due to the much lower number of observations compared to G-spread. As elaborated upon in the model input section of the study, G-spread is a more sophisticated measure of bond performance than yield, for which reason this will be used as the dependent variable.

The average G-spread on 31-12-2019 was 478 bps for the private dataset and 249 for the public. At first glance, the extreme values at maximum and minimum for both datasets could be a cause for concern. However, as some of the bonds in the dataset are highly distressed[12] naturally yields will be extremely high, as they simply capture an immense credit risk. Investors in this case often trade on expectations about the recovery rate, and do not expect full repayment of the bonds principal. However, no high yield bond can reasonably be expected to trade at a spread of minus several thousands of bps. However, several factors can cause such datapoints. One is the fact that callable bonds will often trade slightly above their call premium at negative yield, if the underlying company improve its position relative to the time of issuance. This is due to the fact that restructuring costs are substantial from calling an outstanding bond, meaning that a bond trading at 102,5 with an inbuilt call clause at 102 may not be worth redeeming for the company. However, to the investor, the theoretical yield to worst of a bond trading at 102,5 which could be called immediately at 102 is strictly negative. As such, trades are made based on expectations of whether bonds will be called or not.

Another important factor is illiquidity in the high yield market. As illustrated by the quote at the beginning of this paper, the European high yield market is less developed than the equity markets. Since only institutional investors trade in the market, and often do so over the phone, high yield securities may often have longer periods where no trades are made, and no new bids or asks are published in the market. As such, since yield and spreads are always calculated using the latest midprice, spreads can, on paper, be extreme if markets have changed since last time the security was traded. This effect can be amplified in case a callability provision has kicked in since the security last was traded. Lastly, it is important to note that extreme values are only present in very limited amounts, as illustrated by the following histogram for G-spread in the public dataset:

[12] The minimum midprice of the private dataset is 1,48 which should be compared to Capital Fours rule of thumb stating that credits trading below 85 points are in distressed area (Capital Four)

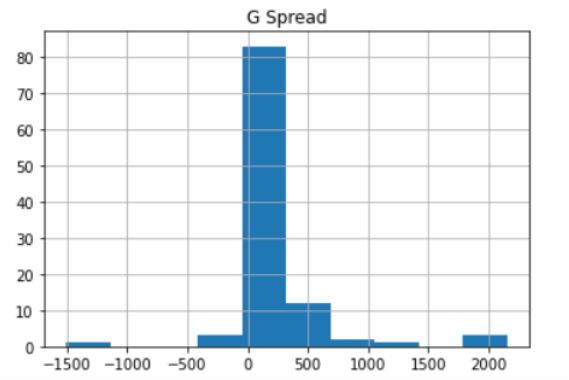


Figure 23: G-spread of EU public dataset

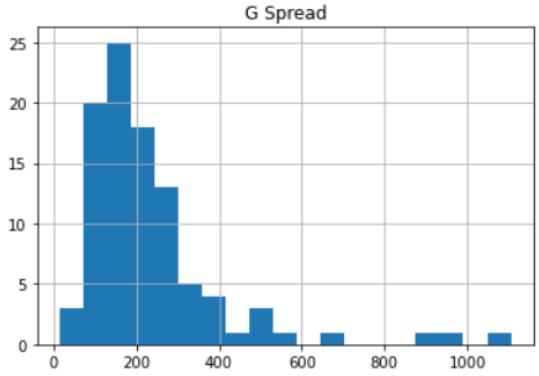


Figure 24: G-spread of EU public dataset after removal of outliers

With four negative values and 1 positive outlier, the bulk of the data falls within common territory for high yield spreads. Removing outliers (figure 23) makes it evident that spreads are close to evenly distributed around the mean of 249 bps, with a positive skew as could be expected.

A more pressing issue is that of negative G-spreads at the time of issue. There is no logical reason, theoretical or empirical, as to why a high yield bond would trade with yield lower than the corresponding government yield curve a single day after being issued. Looking at issue midprice, we can determine which data points should be flagged as outliers. It is common for bond issuances to go to the market with a discount. Printing a bond at price 98 rather than at par is a simple way to offer investors a greater yield, similar to increasing the coupon of the bond. This explains how issue midprice can be several points below par. However, an issue midprice much lower than this would either indicate data flaws (for instance a price at January 1st 2014 when the bond had in fact been trading before this date) or a disastrous issue. We employ the following check to ensure data validity for issue spreads: The 5 datapoints with issue midprice < 97 are checked, and found to relate to issues earlier than 2014, and are consequently removed from the dataset. Additionally, two instances of negative issue G-spread are removed.

With the dependent variable in place, the final step is to combine descriptive bond data, accounting data, data of textual features from bond prospectuses and spreads. Doing so yields a dataset of 176 datapoints for the private dataset, 91 datapoints for the public dataset predicting spreads at 31-12-2019 and 79 datapoints for the public dataset predicting spreads at issue.

7. Analysis – Explaining high yield spreads using linear regression

With all data collected and processed, we begin the actual model setup. Two separate sections will outline the models employed, one focusing on multiple linear regression as a tool to explain credit spread, and another focusing on several machine learning algorithms, used in an attempt to predict spreads on unseen data. We begin by setting up the linear regression.

7.1 Model specification - Linear regression

The initial model applied consists of a multiple linear regression with the goal of determining G-spread as a function of accounting variables, bond specific variables and text analysis variables. The model takes the following form:

$$y_i = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots \beta_n \cdot x_n + u_i$$

Where y is the dependent variable, α is a constant, β_n is a coefficient describing the effect of x_n on the dependent variable and u_i is an error term.

7.2 Evaluation metrics

The simplest evaluation metric of a linear regression is R^2:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where y_i is the true value of the dependent variable for datapoint i, $\hat{y}_i$ is the modelled (predicted) value, and $\hat{y}$ is the mean of the dependent variables is the sample. In other words, R^2 is the fraction of total variance explained by the model. The issue with using R^2 is the fact that adding new explanatory variables will always increase the measure, regardless of whether they hold actual explanatory power.

As the study employs numerous variables, it is important to evaluate models with an appropriate measure. We use adjusted R^2, which is a measure similar to R^2, but with the added effect of punishing the inclusion of independent variables with no explanatory power. Adjusted R^2 is defined as:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

Where n is the number of datapoints in the sample, and p is the number of explanatory variables included in the model (excluding the constant α).

Adjusted R^2 is the commonly accepted measure for accuracy of multiple regression models (Stock & Watson, 2015)

7.3 Model setup

In order to create a setup of the study that allows for comparison between models including text features and models not including text features, we determine the optimal adjusted R^2 models in three steps: First, by using solely rating, secondly by including bond and accounting variables, and lastly, by including text variables.

As such, an initial model is setup based solely on the rating of the bond in question. The full specification of the model looks like the following:

$$Spread_i = \alpha + \beta_1 \cdot Rating_i + u_i$$

Where $Spread_i$ is the dependent variable, G-spread of $bond_i$, $Rating_i$ is the numerical value assigned to the Bloomberg composite rating for $bond_i$, α is the intercept, β_1 is the estimator of the effect of rating on spread and u_i is the error term.

Secondly, we setup a model including bond specific and accounting variables. The full specification of the model looks like the model containing only rating, but instead also contains estimators for the nine accounting ratios: Net leverage, FCF to Net Debt, cash ratio, adjusted EBITDA margin, net income margin, equity cushion, working capital to assets, asset tangibility and sales growth, as well as dummy variables for industry, currency, geography, seniority and callability.

The next step is to remove unwanted features, that is, features with low or no explanatory power. There are several reasons as to why the model could contain more features than optimal. Perhaps a parameter simply does not explain bond spreads on any level of statistical significance, in which case it should be removed. But it could also be that two measures are very correlated, and that the effect of one of the measures on the spread is therefore captured in another measure.

Once the model with the largest value of adjusted R^2 is achieved, the text features are added. This third model is then evaluated against the first and second, and potential text feature variables with no explanatory power is removed.

7.3.1 Assumptions

All estimators are estimated using ordinary least square (OLS) regression, which minimizes the residual sum of squares $\sum_i (y_i - \hat{y}_i)^2$. As the models are estimated using standard linear regression, the details of the OLS estimator will not be explained in any further details. However, it is crucial to be aware of the assumptions behind the estimator, as these are required to be fulfilled for the estimators to be consistent. Below are the most important for the study in question, in accordance with Stock & Watson (2015).

7.3.1.1 Variables must be i.i.d.

The variables in the model must be independently and identically distributed, which implies that the sample must be representative of the overall population. In order to adhere to this assumption, random sampling is often used when selecting the test sample. In this study, data has been drawn

from commercial corporations, and the fact that only some of the initial bonds were successfully linked to the corresponding financial data of the underlying company, either because of data availability or other, could introduce a bias in which only well-functioning companies tracked by data aggregators such as Bloomberg or 9fin make it into the dataset. It is not possible to reject that such a bias may exist, but from the sample split across industry, geography, currency and ranking it seems plausible that the dataset is a good representation of the overall European high yield universe.

7.3.1.2 No perfect multicollinearity

Two independent variables cannot be completely positively or negatively correlated. While the regression software automatically raises an error if this is ever the case, having variables that are very closely correlated can mean that coefficients are not estimated correctly, as the effect that two independent variables has on the dependent variable cannot be separated. Below is the correlation matrix of the accounting variables:

Figure 25: Correlation matrix of quantitative input features, EU private dataset

We observe that no variables are anywhere remotely close to being perfectly correlated, but that the correlation between the three profitability measures are in the range of 0,5 to 0,6 which is understandable. Another logical source of multicollinearity would be between industry variables and accounting ratios, as firms from the same industry generally have similar ranges of profitability, leverage etc. Similarly, we must ensure that the rating assigned to a security is not perfectly correlated with the accounting variables, which may have been used when assigned the

ratio. However, even if industry variable dummies as well as rating scores are included, no alarming correlations are found:

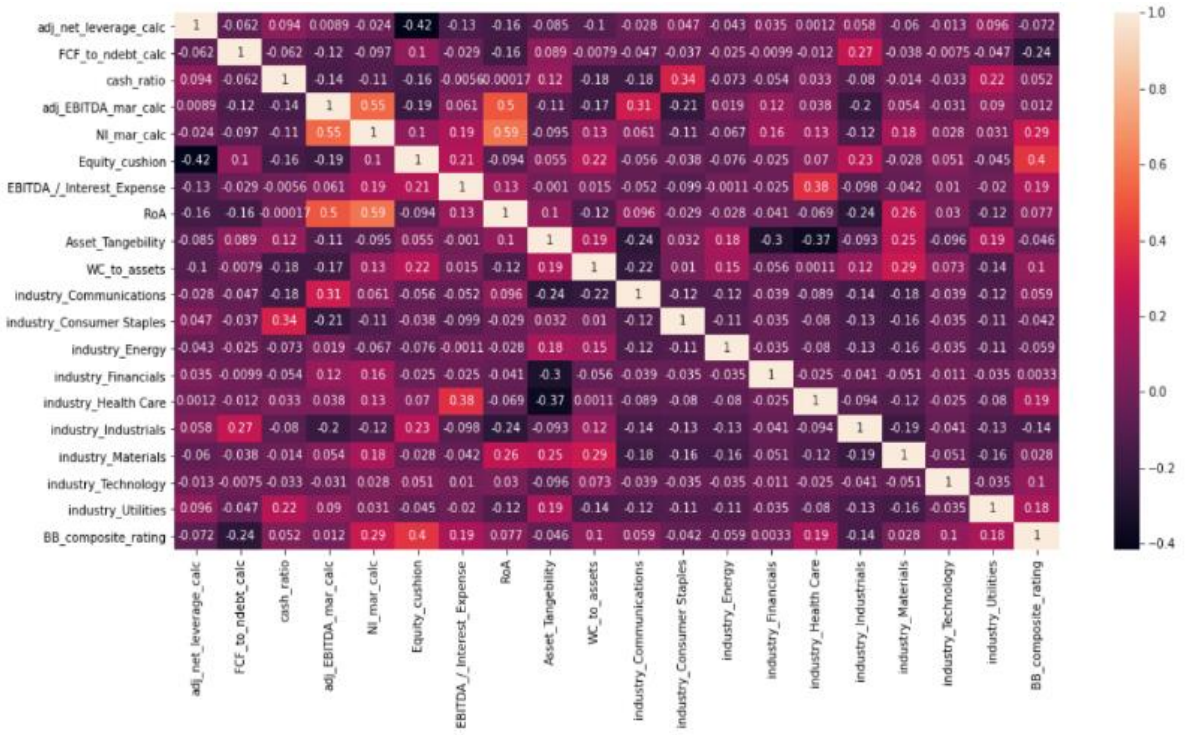


Figure 26: Correlation Matrix of quantitative input features, EU Public dataset

Indeed, the most correlated variables remain the profitability measures discussed above. Interestingly, the credit rating score is most correlated with equity cushion, indicating that rating agencies assign large weight to this measure when assigning credit ratings.

7.3.1.3 The error term u has a conditional mean of 0 and should be uncorrelated with any past, future and present values of the explanatory variables

Meaning that the error term of the model is independent of any other variable and that it should be normally distributed. When this is the case, the data is said to be homoscedastic. If this is not the case, the model estimators will not be consistent, and statistical inference cannot be drawn from any results, although the model is still robust. Ensuring that the dataset is homoscedastic is therefore important for any conclusions drawn from the models, and consequently we employ both a Breush-Pagan and a White test, which are the most commonly used tests used to check for heteroscedasticity (Stock & Watson, 2015). The results are listed in the following table:

Table 8: Breusch Pagan Test and White Test results

Test for heteroscedasticity

Test name	Breusch Pagan test				White test			
Dataset	LM statistic	LM P-value	F-Statistic	F-test P value	LM statistic	LM P-value	F-Statistic	F-test P value
Private	0,8386002	0,9331995	0,2039908	0,9358616	4,0459459	0,9951834	0,2680250	0,9962519
Public, 31-12-2019	2,4579723	0,8731413	0,3886472	0,8844023	7,2180309	0,9999481	0,2010226	0,9999894
Public, Issue	2,6184568	0,8549820	0,4113753	0,8691940	14,9769217	0,9697544	0,4418679	0,9879062

By examining test P values, it is evident that the null hypothesis of heteroscedasticity cannot be rejected, both when evaluating using Lagrange Multiplier (LM) or F statistic. We therefore employ regular non-robust standard errors in order to determine statistical significance of estimators in the model.

7.3.1.4 Finite fourth moment

In any distribution, the fourth moment, or the kurtosis, refer to the relative importance of tail versus body in the distribution. As such, a large fourth moment is caused by observations far from the means, or outliers. Finite fourth moment should therefore be interpreted as a dataset in which large outliers are unlikely. The presence of large outliers can negatively affect estimators of the model, making them inconsistent. As discussed earlier, outliers were removed when reviewing the data and when calculating accounting ratios. Similarly, by observing scatterplots of the variables in question is, it appears that this assumption is met. It is also important to note that simply removing any datapoints on the margin of observation in order to adhere to this assumption would be counterproductive, as this would decrease the level of generalization one can infer from any results. As such, values that are simply larger than usual, such as the case with maximum observations of interest coverage ratios in both the public and private dataset, are left untouched.

8. Results – Linear Regression

Results of the initial model with only rating as an explanatory variable for the public dataset:

OLS Regression Results

Dep. Variable:	Gspread_31_12_19	R-squared:	0.160
Model:	OLS	Adj. R-squared:	0.150
Method:	Least Squares	F-statistic:	16.89
Date:	Fri, 08 May 2020	Prob (F-statistic):	8.78e-05
Time:	21:30:34	Log-Likelihood:	-647.67
No. Observations:	91	AIC:	1299.
Df Residuals:	89	BIC:	1304.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	890.4489	166.030	5.363	0.000	560.550	1220.348
BB_composite_rating	-61.2692	14.907	-4.110	0.000	-90.888	-31.650

Omnibus:	61.504	Durbin-Watson:	1.797
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1589.327
Skew:	-1.384	Prob(JB):	0.00
Kurtosis:	23.286	Cond. No.	58.9

Figure 27: EU public, linear regression results, rating only

Note that rating is a significant explainer of G-spread, and that adjusted R^2 is 0,15.

Results of the same regression, adding bond descriptive and accounting variables (step 2):

OLS Regression Results

Dep. Variable:	Gspread_31_12_19	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.322
Method:	Least Squares	F-statistic:	2.424
Date:	Mon, 11 May 2020	Prob (F-statistic):	0.00178
Time:	13:08:35	Log-Likelihood:	-619.46
No. Observations:	91	AIC:	1301.
Df Residuals:	60	BIC:	1379.
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	978.4635	308.270	3.174	0.002	361.832	1595.095
BB_composite_rating	-54.2776	18.756	-2.894	0.005	-91.795	-16.760
ccy_GBp	18.3711	112.999	0.163	0.871	-207.661	244.404
ccy_USD	-48.3983	139.316	-0.347	0.730	-327.071	230.274
ccy_other	32.1679	140.586	0.229	0.820	-249.046	313.382
geo_Other	-15.2025	124.021	-0.123	0.903	-263.282	232.877
geo_South_EU	-292.4783	121.773	-2.402	0.019	-536.060	-48.897
geo_US	-29.3231	169.828	-0.173	0.863	-369.029	310.383
industry_Consumer Discretionary	23.6529	94.348	0.251	0.803	-165.071	212.377
industry_Communications	77.3826	99.169	0.780	0.438	-120.984	275.749
industry_Consumer Staples	-5.4699	110.977	-0.049	0.961	-227.456	216.517
industry_Energy	90.4730	116.202	0.779	0.439	-141.966	322.911
industry_Financials	56.3841	276.008	0.204	0.839	-495.714	608.482
industry_Health Care	134.5388	142.561	0.944	0.349	-150.626	419.704
industry_Industrials	109.8647	102.500	1.072	0.288	-95.165	314.894
industry_Materials	71.0775	98.652	0.720	0.474	-126.255	268.410
industry_Technology	139.8220	280.159	0.499	0.620	-420.580	700.224
industry_Utilities	280.7380	141.569	1.983	0.052	-2.442	563.918
Sales_3Y_CAGR	-3.7278	2.902	-1.285	0.204	-9.533	2.077
Issue_size	1.975e-08	2.07e-08	0.952	0.345	-2.17e-08	6.12e-08
Callable_Y	-61.1836	95.137	-0.643	0.523	-251.486	129.119
Seniority_SNDB	-25.3438	116.968	-0.217	0.829	-259.314	208.627
adj_net_leverage_calc	7.2979	9.139	0.799	0.428	-10.982	25.578
FCF_to_ndebt_calc	45.4479	8.193	5.547	0.000	29.060	61.836
cash_ratio	-3.4888	2.544	-1.371	0.175	-8.578	1.600
adj_EBITDA_mar_calc	48.9349	255.093	0.192	0.849	-461.327	559.197
NI_mar_calc	-697.8978	576.416	-1.211	0.231	-1850.901	455.106
Equity_cushion	-27.0304	117.213	-0.231	0.818	-261.491	207.430
EBITDA_/_Interest_Expense	-1.7323	1.913	-0.905	0.369	-5.559	2.095
RoA	541.7756	777.932	0.696	0.489	-1014.320	2097.871
Asset_Tangebility	-227.5932	239.937	-0.949	0.347	-707.540	252.353
WC_to_assets	631.1559	323.477	1.951	0.056	-15.894	1278.205

Omnibus:	85.203	Durbin-Watson:	1.912
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1186.367
Skew:	-2.736	Prob(JB):	2.42e-258
Kurtosis:	19.821	Cond. No.	6.89e+15

Figure 28: EU public, linear regression results, rating and quantitative features

Adjusted R^2 is now 0,322. After feature elimination, the following regression is produced (modified step 2):

OLS Regression Results

Dep. Variable:	Gspread_31_12_19	R-squared:	0.538
Model:	OLS	Adj. R-squared:	0.388
Method:	Least Squares	F-statistic:	3.593
Date:	Mon, 11 May 2020	Prob (F-statistic):	2.68e-05
Time:	13:08:50	Log-Likelihood:	-620.48
No. Observations:	91	AIC:	1287.
Df Residuals:	68	BIC:	1345.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	913.8082	206.869	4.417	0.000	501.007	1326.610
BB_composite_rating	-53.8865	15.064	-3.577	0.001	-83.946	-23.827
geo_Other	33.9263	82.452	0.411	0.682	-130.603	198.456
geo_South_EU	-308.7549	105.584	-2.924	0.005	-519.445	-98.065
geo_US	-94.2369	120.173	-0.784	0.436	-334.038	145.564
industry_Consumer Discretionary	22.6366	83.063	0.273	0.786	-143.113	188.387
industry_Communications	84.6935	86.383	0.980	0.330	-87.680	257.067
industry_Consumer Staples	-5.0992	100.312	-0.051	0.960	-205.269	195.071
industry_Energy	109.1842	106.128	1.029	0.307	-102.591	320.959
industry_Financials	18.5013	251.162	0.074	0.941	-482.684	519.686
industry_Health Care	101.2218	130.540	0.775	0.441	-159.267	361.711
industry_Industrials	79.5789	86.468	0.920	0.361	-92.965	252.123
industry_Materials	82.0987	87.088	0.943	0.349	-91.682	255.880
industry_Technology	150.1900	260.228	0.577	0.566	-369.087	669.467
industry_Utilities	270.8024	115.866	2.337	0.022	39.596	502.009
Sales_3Y_CAGR	-3.8703	2.640	-1.466	0.147	-9.137	1.397
Issue_size	1.732e-08	1.67e-08	1.035	0.304	-1.61e-08	5.07e-08
adj_net_leverage_calc	7.1365	7.396	0.965	0.338	-7.623	21.896
FCF_to_ndebt_calc	44.5026	7.647	5.819	0.000	29.243	59.762
cash_ratio	-3.4095	2.250	-1.515	0.134	-7.900	1.081
NI_mar_calc	-380.2520	363.787	-1.045	0.300	-1106.179	345.675
EBITDA_/_Interest_Expense	-1.5195	1.674	-0.908	0.367	-4.860	1.821
Asset_Tangebility	-166.1972	205.630	-0.808	0.422	-576.526	244.131
WC_to_assets	492.8403	274.166	1.798	0.077	-54.249	1039.930

Omnibus:	83.508	Durbin-Watson:	1.880
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1169.729
Skew:	-2.648	Prob(JB):	9.92e-255
Kurtosis:	19.747	Cond. No.	1.39e+25

Figure 29: EU public at 31/12/2019, linear regression results, rating and best quantitative features

With adjusted R^2 of 0,388 this represents the best regression before adding textual features. Below is the result of the same regression, with textual features added as explanatory variables (step 3):

OLS Regression Results

Dep. Variable:	Gspread_31_12_19	R-squared:	0.608
Model:	OLS	Adj. R-squared:	0.402
Method:	Least Squares	F-statistic:	2.949
Date:	Mon, 11 May 2020	Prob (F-statistic):	0.000175
Time:	13:09:03	Log-Likelihood:	-612.99
No. Observations:	91	AIC:	1290.
Df Residuals:	59	BIC:	1370.
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1839.6468	589.880	3.119	0.003	659.300	3019.993
BB_composite_rating	-75.4509	18.186	-4.149	0.000	-111.842	-39.060
geo_Other	14.5741	103.216	0.141	0.888	-191.961	221.109
geo_South_EU	-233.1617	116.118	-2.008	0.049	-465.513	-0.811
geo_US	-155.0562	130.621	-1.187	0.240	-416.428	106.316
industry_Consumer Discretionary	50.3536	92.324	0.545	0.588	-134.387	235.094
industry_Communications	205.6171	110.407	1.862	0.068	-15.306	426.541
industry_Consumer Staples	85.1425	114.770	0.742	0.461	-144.512	314.797
industry_Energy	152.9312	126.544	1.209	0.232	-100.282	406.145
industry_Financials	246.8357	279.017	0.885	0.380	-311.476	805.148
industry_Health Care	189.7459	143.583	1.322	0.191	-97.563	477.055
industry_Industrials	125.9498	94.836	1.328	0.189	-63.817	315.717
industry_Materials	155.7843	100.246	1.554	0.126	-44.808	356.376
industry_Technology	235.4101	277.041	0.850	0.399	-318.947	789.767
industry_Utilities	391.8877	147.346	2.660	0.010	97.049	686.726
Sales_3Y_CAGR	-2.9105	3.220	-0.904	0.370	-9.355	3.534
Issue_size	1.229e-08	1.89e-08	0.652	0.517	-2.55e-08	5e-08
adj_net_leverage_calc	6.3725	7.703	0.827	0.411	-9.041	21.786
FCF_to_ndebt_calc	40.4864	8.299	4.878	0.000	23.879	57.093
cash_ratio	-1.5922	2.473	-0.644	0.522	-6.541	3.357
NI_mar_calc	-514.1924	390.923	-1.315	0.193	-1296.428	268.044
EBITDA_/_Interest_Expense	-0.6606	1.795	-0.368	0.714	-4.252	2.930
Asset_Tangebility	-219.1901	218.440	-1.003	0.320	-656.288	217.908
WC_to_assets	491.2589	341.728	1.438	0.156	-192.537	1175.055
topic_1_weight	638.0465	194.194	3.286	0.002	249.466	1026.627
topic_2_weight	544.5476	223.579	2.436	0.018	97.166	991.929
topic_3_weight	657.0470	199.011	3.302	0.002	258.827	1055.267
neg_share	-3987.2768	4596.784	-0.867	0.389	-1.32e+04	5210.867
pos_share	5656.9379	1.46e+04	0.389	0.699	-2.35e+04	3.48e+04
uncer_share	1546.0417	5803.705	0.266	0.791	-1.01e+04	1.32e+04
lit_share	-662.6189	5837.775	-0.114	0.910	-1.23e+04	1.1e+04
tot_word	-0.0112	0.008	-1.420	0.161	-0.027	0.005
sentiment_polarity	-3052.7717	2859.320	-1.068	0.290	-8774.257	2668.714
subjectivity	-2573.0208	1847.539	-1.393	0.169	-6269.938	1123.897

Omnibus:	52.215	Durbin-Watson:	1.861
Prob(Omnibus):	0.000	Jarque-Bera (JB):	316.604
Skew:	-1.640	Prob(JB):	1.78e-69
Kurtosis:	11.529	Cond. No.	1.08e+16

Figure 30: EU public at 31/12/2019, linear regression results, rating, best quantitative features, and text features

Adjusted R^2 is 0,402 meaning that explanatory power increased. The final regression is an identical regression with insignificant parameters removed (modified step 3):

OLS Regression Results

Dep. Variable:	Gspread_31_12_19	R-squared:	0.602
Model:	OLS	Adj. R-squared:	0.432
Method:	Least Squares	F-statistic:	3.534
Date:	Mon, 11 May 2020	Prob (F-statistic):	1.91e-05
Time:	13:10:08	Log-Likelihood:	-613.62
No. Observations:	91	AIC:	1283.
Df Residuals:	63	BIC:	1354.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1758.2923	527.742	3.332	0.001	703.684	2812.900
BB_composite_rating	-74.1776	17.524	-4.233	0.000	-109.198	-39.158
geo_Other	1.5643	88.377	0.018	0.986	-175.043	178.171
geo_South_EU	-240.3438	112.453	-2.137	0.036	-465.063	-15.625
geo_US	-131.6237	121.065	-1.087	0.281	-373.553	110.306
industry_Consumer Discretionary	44.8504	83.813	0.535	0.594	-122.636	212.337
industry_Communications	211.6432	104.338	2.028	0.047	3.141	420.145
industry_Consumer Staples	96.9210	106.185	0.913	0.365	-115.272	309.114
industry_Energy	169.0348	117.306	1.441	0.155	-65.382	403.452
industry_Financials	235.0249	261.685	0.898	0.373	-287.911	757.961
industry_Health Care	168.2413	137.233	1.226	0.225	-105.996	442.479
industry_Industrials	134.5052	90.844	1.481	0.144	-47.032	316.043
industry_Materials	142.8946	95.572	1.495	0.140	-48.091	333.880
industry_Technology	203.8045	265.115	0.769	0.445	-325.985	733.594
industry_Utilities	351.3723	135.626	2.591	0.012	80.344	622.400
Sales_3Y_CAGR	-2.1588	2.881	-0.749	0.456	-7.916	3.599
Issue_size	7.734e-09	1.7e-08	0.455	0.650	-2.62e-08	4.17e-08
adj_net_leverage_calc	5.6210	7.419	0.758	0.452	-9.206	20.448
FCF_to_ndebt_calc	38.9321	7.858	4.955	0.000	23.230	54.635
cash_ratio	-1.2392	2.346	-0.528	0.599	-5.928	3.450
NI_mar_calc	-503.6965	376.331	-1.338	0.186	-1255.734	248.341
EBITDA_/_Interest_Expense	-0.8158	1.724	-0.473	0.638	-4.260	2.628
Asset_Tangebility	-177.1773	206.467	-0.858	0.394	-589.769	235.415
WC_to_assets	414.6213	301.693	1.374	0.174	-188.263	1017.506
tot_word	-0.0139	0.007	-2.081	0.042	-0.027	-0.001
sentiment_polarity	-1669.9865	2297.218	-0.727	0.470	-6260.610	2920.637
subjectivity	-2660.9594	1779.905	-1.495	0.140	-6217.816	895.898
topic_1_weight	617.5100	174.288	3.543	0.001	269.224	965.796
topic_2_weight	506.8275	194.072	2.612	0.011	119.005	894.650
topic_3_weight	633.9549	184.626	3.434	0.001	265.009	1002.901

Omnibus:	53.994	Durbin-Watson:	1.859
Prob(Omnibus):	0.000	Jarque-Bera (JB):	343.742
Skew:	-1.693	Prob(JB):	2.28e-75
Kurtosis:	11.899	Cond. No.	1.08e+16

Figure 31: Figure 29: EU public at 31/12/2019, linear regression results, rating, best quantitative features, and best text features

With adjusted R^2 of 0,432 the model has the largest explanatory power of the models tested.

Below is a summary of model results for all datasets:

Linear Regression

Dataset	Private				Public - 31-12-2019				Public - Issue			
Model setup	R^2	Adjusted R^2	F statistic	F-test P value	R^2	Adjusted R^2	F statistic	F-test P value	R^2	Adjusted R^2	F statistic	F-test P value
Rating	0,2310000	0,1100000	12,4660325	0,0000910	0,1600000	0,1500000	16,8900000	0,0000876	0,5950000	**0,5830000**	14,7200973	0,0054515
Step 2	0,3120000	0,2710000	2,1771049	0,0011339	0,5480000	0,3220000	2,4240000	0,0017800	0,6910000	0,4110000	0,9499377	0,0064931
Modified step 2	0,3260000	0,3190000	0,4104454	0,0000760	0,5380000	0,3880000	3,5930000	0,0000268	0,6570000	0,4520000	1,0692894	0,0003677
Step 3	0,6100000	0,4100000	0,6555799	0,0020842	0,6080000	0,4020000	2,9490000	0,0001750	0,7210000	0,4320000	1,9703108	0,0074435
Modified step 3	0,6020000	**0,5920000**	3,9613950	0,0000039	0,6020000	**0,4320000**	3,3534000	0,0000198	0,6640000	0,4750000	2,0676946	0,0000639

Figure 32: Summary of results of linear regression, all datasets

8.1 Discussion of results

As could be expected given the levels of variables, all models have F-test P values indicating statistically significant results at less than 1% confidence level. Interestingly, prediction of spreads at 31-12-2019 displays similar characteristics for private and public firms, with the most noticeable difference being the level of information captured in textual features of bond prospectuses. For private firms, the explanatory power increases from 0,319 to 0,592 (a difference of 0,273) when moving from the most optimal model without textual features (Modified step 2) to the most optimal model with textual features (Modified step 3). For public firms, the difference is only 0,044 indicating that the benefit of text analysis in explaining bond spreads is lower. This finding seems intuitively in line with theory, as publicly traded companies have a constant flow of information to investors in the form of earnings updates, management calls, capital market days etc. As such, the information contained in the original bond prospectus is updated and refined. For private firms, such information is available to a much lower degree, meaning that the bond prospectus is still an important source of information when setting expectations about future performance. This effect would also explain why a larger proportion of total variance can be explained for private firms than for public (0,592 versus 0,432).

Surprisingly, neither accounting variables nor textual features improves the explanatory power of the model predicting spreads as issue for the public dataset. As mentioned, public high yield bonds will often be bond issued with investment grade ratings, so called fallen angels. One of the major distinctions between analysis of corporate investment grade bonds versus analysis of corporate high yield bonds is the importance of credit risk. The rating at issue, which for many of the securities are BBB+ or higher, signals very low risk of default to investors, meaning that credit risk explains a smaller proportion of the spread, as shown by Longstaff et al (2005) and Huang & Huang (2003). Consequently, the interest rate and liquidity become a more central question, which accounting variables and textual features do not provide information on. But even for public bonds originally issued as high yield, it seems reasonable that rating has more explanatory power closer to the time of issue, as this is the time where a security is assigned its rating[13]. This effectively implies that the rating better captures financial information, such as accounting variables, for which reason they, along with textual features, will have less additional explanatory power in explaining spread variance.

9. Analysis - Machine Learning framework for predicting spreads

With multiple linear regression having shown a statistically significant explanatory power of textual features from bond prospectuses on bond yield spreads, the next section of the paper will investigate the ability to predict such spreads using the same model inputs, but applying various machine learning algorithms. The multiple linear regression allows for an academic understand

[13] Ratings are updated regularly, but usually only following severe alterations to the financial health of the underlying company

of relationships between variables and is easily interpretable. However, more advanced methodologies can account for relationships between variables not properly captured using linear regression, and these methodologies can therefore and as such, we now setup a framework similar to Tao & Deokar (2015) and Deokar (2018) studies on equity IPO underpricing and apply it to the universe of European high yield[14].

9.1 Train / test split

A main difference between the regular linear regression from the previous section and the setup of the machine learning models is the way the model performances are evaluated. The linear regression model assumes that the dataset is a random subsample of the overall distribution, and thus representative of the overall distribution within regular standard errors. The model was fitted on the entire dataset and evaluated on the basis of Adjusted-R^2, to control for the problem of overfitting.

For the evaluation of the machine learning models, we adopt the best practice method of evaluation within the field of machine learning of splitting the dataset into two sub-set: a training-set and a testing-set (Daumé, 2017). Splitting the data into a training-set and a testing-set allows for an iterative process of finetuning the parameters of the different models and different model configurations to increase the performance of the predictive model, while still being able to evaluate the final performance of the model on the test-set which is data that is unknown for the model, meaning that the model has not been fitted on it. By testing on an unseen test set, the evaluation problem of overfitting the data is solved as well. If the model overfits the training-data, it will perform significantly worse on the unseen test data, as this will just follow the underlying data distribution, and not the specific structures and outliers found in the particular sample used to construct the model (Daumé, 2017).

9.2 Evaluation metrics

As a result, every model is evaluated on how well it predicts the datapoints in the test-set. This makes it possible to compare and evaluate different models across a set of evaluation metrics or loss-functions (a measure of how wrong the model has predicted the data). The Y-values we are trying to predict are risk spreads, which is a continuous variable, and thus the evaluation metrics must be effective in evaluating the performance of regression on a continuous value. For evaluation of the models, we apply the following four evaluation metrics: Mean Absolute Errors (MAE), Root Mean Squared Errors (RMSE), R-Squared (R^2), and Explained Variance. Each of these do in essence indicate how well the model have predicted the dataset, but they each provide different nuances of information about how well the model have predicted the data points, while also being widely adopted metrics for model evaluation (Casella et al, 2017). Mean Absolute Error

[14] Both studies referenced are set up as a classification problem, predicting underpricing as a binary yes/no variable, but as the independent variable in this study is a continuous variable, we instead setup a regression problem.

is a measure of the average error between paired observations of predicted and actual values. Hence it gives an indication on the average margin of error on predictions. To find MAE one has to calculate in absolute terms how big is the difference between each predicted value of the model and the actual model, and then divide by the total number of observations (Stock & Watson, 2015):

$$MAE = \sum_{i=1}^{n} \frac{abs(y_i - \hat{y_i})}{N}$$

Root Mean Squared Error (RMSE), is similar to MAE, as it also expresses the average model prediction error. But instead of simply taking the absolute value of the pairwise errors, RMSE square the errors before taking the average. As a result, it gives higher weight to large error, and thus better for outlier detection and to detect large errors caused by the model (Stock & Watson, 2015).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{N}}$$

As defined in the section on linear regression, R^2, also known as the coefficient of determination, is the proportion of the of the variance in the dependent variable that can be predicted from the independent variables (Stock & Watson, 2015). It is thus, different from the two other metrics, as instead of measuring the average error it describes how well the model captures the underlying variance in the dependent variable. If the dependent variable has very high variance, but the model also predicts the variance very well, it will have high R^2-score, but the model can still return high MAE and RMSE values, due to the underlying high variance in the dependent variable. This makes the R^2 better for comparison across datasets.

Explained variance is a score that, similar to R^2, captures how much of the overall variance in the dependent variable is explained by the model. It is calculated as variance of the errors when predicting the dependent variable, divided by the overall variance of the dependent variable (Griez et al, 2016):

$$Explained\ Variance\ = 1 - \frac{Var(y_i - \hat{y_i})}{Var(y)}$$

For splitting the data into a training and test subset, we adopt the industry norm of an 80/20 split, with 80% of the data split into the training set and 20% into the test set (Daumé, 2017). This leaves a large part of the data to optimize the algorithms, while still leaving a substantial dataset for evaluation afterwards. We split the data randomly, to make sure the test set is representative

of the overall distribution in the data. To carry out the random split we used the training/test split algorithm from the Sci-Kit learn python library[15].

With the dataset split and evaluation metrics defined, the next step is to setup a framework of machine learning regression algorithms. Machine learning regression can be performed with a variety of different algorithms, but, as stated by Macready & Wolpert's theorem as early as 1997, there is no such thing as a free lunch in machine learning. There will be inherent trade-offs for each model and algorithm, and no model is universally better. The performance of different machine-learning algorithms often depends on the size and structure of the underlying data (Daumé, 2017). These structures are often hidden, in the sense that it is not structures easily picked up by the human mind. As a result, the correct choice of algorithm and model setup will often remain unclear until the models have been tested and evaluated (Géron, 2017). We setup a framework of four different machine learning models. We setup a linear Ridge regression model, a Random Forest model, and a Support Vector Machine (SVM) model, which are all very prominent and widely used for regression problems (Géron, 2017). Furthermore, we setup an Ensemble model, which is a combination of the regression predictors of the three other models, combined into a single regressor. We also originally wanted to set-up and train a Neural Network in the form of a Multi-Layer Perceptron regressor. However, due to the size of our final dataset, we did not have enough data to properly train such a Neural Network, and thus discarded that model. But if the test of this paper is replicated with a larger training-set there would be potential of also setting up a Neural Network, as they are highly effective of modelling highly complex, not linear relationship, which can be expected to be found within NLP analysis (Norvig & Russel, 2010).

All machine learning algorithms contain parameters that are manually set before the training of the models begin. To distinguish them from parameters affected by the data, they are named hyper-parameters. As they can greatly affect the performance of the models, a proper method for fine tuning the hyper-parameters for each model needs to be implemented. For fine tuning each of the models we use grid-search with cross-validation. Grid-search is a pragmatic yet effective way to tune the hyper parameters of a model (Géron, 2017). To conduct a grid-search we fit a model multiple times with different configurations of each hyper-parameter relevant to the model. The result of each pairwise configuration of the hyper-parameters is then laid out in a grid. The configurations with the best performance will then be used as the final model. To evaluate the results of the hyper-parameter configurations we will use the R^2 measure, as this is the one which provide information about the model performance, which is best compared across models. Evaluation of the hyper-parameters will be done using 5-fold cross validation on the training set. With 5-fold cross validation, the training set is divided into five different sub-sets or 'folds'. The model is then trained five times using four of the five folds as a training data and the fifth fold as test data. The average R^2 – score across all five trainings is then given as the R^2 score for the given

[15] The library employs a stochastic shuffler to ensure a completely random split of the data

configuration of the model's hyper-parameters. This way, the models will still not have been tested on the final test set before the final evaluation of the models. As this is a computationally highly requiring task (Géron, 2017) the intervals in the grid are set with reasonable intervals, meaning that hyperparameters applied are not guaranteed to be absolutely optimal.

The following section will lay out our setup of each of the four machine learning models, followed by the final choice of features for the models, which will be determined through recursive feature elimination, and lastly a review of the final model setup and full analytical pipeline.

9.3 Linear Ridge Regression model

The first machine learning regression model we set up is a Linear Ridge Regression model. Like the normal linear regression model from the previous section, Linear Ridge regression, is a linear regression model, with an intercept and variable parameters which is optimized through the Ordinary Least Squares method (OLS) (Casella et al, 2017; Stock & Watson, 2015). OLS will fit the regression parameters that leave the lowest residual sum of squares for the data that it is fitted on. In other words, it fit the training data as well as it can with linear parameters only (Casella et al, 2017). However, the goal of these machine learning models, is not to explain the training set as well as possible, but rather to predict the unseen test sets as accurately as possible. Normal OLS linear regression has a high risk of overfitting the training data, as it tries to minimize the residual sum of squares to a global minimum as the only loss function (Casella, et al, 2017). More specifically it is an unbiased model. Ridge regression can reduce the risk of overfitting by adding bias to the model. In short, bias is related to a model failing to fit the training data as perfectly as without bias. However, it does so by making the predictive model more general, which reduces overfitting and thus makes it better able to fit and predict the test data. This is also known as reducing model variance (Casella et al, 2017). Ridge regression does this by introducing a Lambda parameter to each coefficient for the OLS optimization problem. When solving the OLS problem for ridge regression, each coefficient is fitted not only to reduce the Residual Sum of Squares, but also the coefficients timed the Lambda parameter. Thus, the optimization problem for ridge regression is given as:

$$\hat{\beta}^{ridge} = argmin \, \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

Where B is a vector of all the coefficients and X is a vector of all the independent variable inputs. The minimization problem will thus not fit the parameter coefficient β that will minimize the Residual Sum of Squares, but one that takes λ into account as well. As a result, the ridge regression coefficients will be less sensitive to the training data, which reduces overfitting and makes a more generalized model. The higher the λ, the lower the sensitivity of the coefficients to the training data (Casella et al, 2017). However, λ needs to be set manually. If λ is too high, the model will

underfit the training data and thus also perform poorly on the testing data, if it is set too low it will overfit the training data, which will also affect the prediction of the testing data negatively.

To choose the right value for Lambda (λ) we ran a grid-search on the training data with the ridge regression model and ended up with a λ of 20,

Table 9: Grid search for best lambda value

R2 scores	Lambda								
Ridge	0,1	0,5	1	2	5	10	20	50	100
R2	0,551	0,554	0,559	0,568	0,573	0,578	**0,583**	0,574	0,562

To fit the Linear Ridge Regression model, we used the Sci-Kit Learn python library module *Ridge Regression* (Sci-Kit Learn Documentation, 2020). This library implements an algorithm that fits a Ridge Regression models through the process described above.

9.4 Random Forest

The second machine learning model we set up is a Random Forest model. A Random Forest is an ensemble of individual decision tree models and is considered to be one of the most accurate learning algorithms available (Géron, 2017). Furthermore, it is quite easy to interpret and derive insights from. But most importantly, it can handle a lot of input variables, even in the thousands, and solve for complex non-linear relationships, which makes it an ideal learning algorithm for prediction using NLP input (Murphy, 2012). Lastly, it is also a great algorithm for discovering and interpreting what features are important for predicting the independent variable. The only main downside to Random Forest learning models, are the high computational power needed, and its tendency to sometime overfit the training data set (Murphy, 2012).

As mentioned, the Random Forest is constructed through the setup of several individual decision trees. A decision tree regressor works by giving a broad guess of the independent variable based on an initial split at a value of one of the features, known as a node. It then further narrows the guess into more specific ranges as it walks down through the tree with more splits, until it can make a final prediction. The decision tree is constructed through recursive partitioning. It starts from the root node, which is the first feature split known as the parent node, which can be true or false, which then split into two child nodes. These nodes can then be further split into two child nodes and become parent nodes themselves (Géron, 2017). Each node and split are constructed from the root node where the data is split on the feature that results in the highest information gain (IG). This is an iterative process that continues until the information gain is maximized (Géron, 2017).

$$IG(D_p, f) = I(D_p) - \left(\frac{N_{left}}{N_p} I(D_{left}) + \frac{N_{right}}{N_p} I(D_{right})\right)$$

Where f is the feature to perform the split on, D_p, D_{left}, and D_{right} are the total number of samples at the parent, left and right node. I is the impurity measure, which is a measure of how impure the guess made at the node level is. This measure differs based on the task you want the tree to fit. For classification problems, the two most common measures of impurity is Gini and Entropy (Géron, 2017). However, for our model the task is a regression problem, and the two most common impurity measures for regression problems Mean Square Error (MSE) and Mean Absolute Error (MAE). We fit the trees for the forest using weighted MSE, calculated as:

$$MSE(t) = \frac{1}{N_t} \sum_{i \, \epsilon D_t} \left(y^i - \hat{y}_t\right)^2$$

Where t is the node, N_t is the number of training samples at node t, D_t is the data subset at t. This is a greedy process, which means that at every node level a split is adopted, which maximizes the information gain at that node level, even if that means that it will result in a worse data split further down the nodes of the tree (Géron, 2017).

A problem with singular Decision Tree models is that the above optimization is almost to effective. It is very sensitive to the data it is trained on. This leads to two problems. First the Decision Tree predictions will be sensitive to the training / test split and predictions can be quite different depending on the underlying data used for training. Secondly, individual Decision Tree have a high risk of overfitting the training data (Daumé, 2017). To overcome this problem, we train a Random Forest. As an Ensemble, a Random Forest is a meta-estimator which is a combination of many Decision Tree models. To train a Random Forest several individual Decision Trees are trained in parallel and then aggregated by taking the mean of the predicted values of each individual tree as the predicted value for the Random Forest (Géron, 2017). To train several different Decision Tree models, without letting any of them see our testing data, we train each decision tree model on a subset of the training data using Bootstrap Aggregation (Bagging). With Bagging as the sampling technique, new datasets are produced for each tree by drawing data samples from the original training set until a new dataset of size n is created. Drawing data samples with replacement ensures that the re-sampling is truly random (Daumé, 2017). When drawing with replacement, the likely hood that a sample is not included in the draw is $1 - \frac{1}{n}$. When repeated n times the likelihood of a sample not being drawn is ($\left(1 - \frac{1}{n}\right)^n$ as n gets high this moves towards $1 - \frac{1}{e}$. Which means for large datasets, each Decision Tree in the Random Forest will contain approximately 63% of the samples for each predictor (Géron, 2017). As our biggest training set will contain 140 data points, drawing with replacement leaves each tree with subsets containing $\left(1 - \frac{1}{140}\right)^{140} \approx$ 40% of the samples will be used for each individual tree.

As with the Ridge Regression, a Random Forest model has parameters that need to be configured manually. Even though the Random Forest reduce the overfitting of the individual Decision Trees,

if the overfitting created by each single tree is large, there will still be an element of overfitting left in the Random Forest model. To reduce overfitting, it is common to put in a stopping condition, which stops the building of the tree at some node level prior to what would be the maximum information gain, if IG has not been maximized at that level yet. The most prominently used stopping condition is to set a max node depth of each individual tree (Géron, 2017). As with the determination of the Lambda level for the Ridge Regression, setting the right max node depth a Low maximum node depth will add a lot of bias which can potentially reduce variance (Casella et al, 2017). The other parameter is the number of estimators used to create the full Random Forest. In general, following the law of large numbers, more individual estimators will yield better result, until the max level of possible estimators equals the number of sample datapoints (Géron, 2017). But at some level, the added effect of adding further estimators will be insignificant and just add further unnecessary computational requirements to the model.

We fit the model on our dataset using the Sci-Kit Learn *Random Forest Regressor* library module, which trains a Random Forest based on the method described above and with weighted MSE as the impurity measure for the nodes (Sci-Kit Learn Documentation, 2020). To choose the right parameters of *max depth* and *n estimators*, we ran a grid search, based on cross-validation:

Table 10: Grid search results on Max Depth and Number of estimators for Random Forest

R2 scores							
RF	Max Depth						
N estimators	3	5	10	15	20	30	50
2	0,503	0,568	0,584	0,569	0,546	0,519	0,489
5	0,526	0,591	0,607	0,593	0,569	0,542	0,512
10	0,547	0,611	0,627	0,613	0,589	0,562	0,532
15	0,565	0,629	0,645	0,631	0,607	0,581	0,550
20	0,576	0,640	**0,656**	0,642	0,618	0,591	0,561
30	0,580	0,644	0,660	0,646	0,622	0,595	0,565
50	0,586	0,650	0,666	0,652	0,628	0,601	0,571

The grid-search resulted in a Random Forest model with a max-depth of 10 nodes, which was the configuration that left the highest R^2. For the number for estimators we chose n_estimators = 20 as after this level the improvements to R^2 was very minor and further estimators would therefore make the model unnecessarily heavy.

9.5 Support Vector Machine Regression

The third Machine Learning regression model employed is a Support Vector Machine Regression (SVR). In linear regression (or polynomial regression) the aim is to fit a line that fits the data the best. Support Vector Regression functions similarly, but introduces a margin of error ε, error being defined as the difference between predicted and actual value of the independent variable. SVR regression then fits the line that will have the smallest breaches of the tube (the band of ε around

the line) possible (Murphy, 2012). As a result, SVR is not affected by errors that are smaller than the accepted margin of error ε, and only the error values larger than the ε value will be contributing to the final cost function. This makes SVR very memory efficient, which allows it to both run quick on larger datasets, but also to model more complex relationship in the data, beyond simple linear relationships (Murphy, 2012). Fitting a SVR model differs from OLS regression where you minimize the squared errors:

$$Argmin \, \|y - XW\|_2^2$$

Where B is a vector of all the coefficients and X is a vector of all the independent variable inputs. Instead an SVR model will be fitted by minimizing the l2 norm of the vector of coefficients, constraint to the maximum accepted margin of error ε:

$$Argmin \frac{1}{2} \|W\|^2$$

Subject to:

$$|y_i - w_i x_i| \leq \varepsilon$$

However, for almost all data, it is impossible to fit a line that will have all data points within the accepted margin of error without setting a very large accepted margin of error (Géron, 2017). An SVR therefore has to allow some data points to fall outside of the constraint. This introduces a hyper-parameter to the model: level of slack, denoted as C, which is how harshly the SVR model will penalize breaches of the accepted margin of error. The fitting optimization problem is therefore:

$$Argmin \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{n} |\xi_i|$$

Subject to:

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i|$$

Where ξ_i, is the deviation from the margin of error ε , and C is the sensitivity to the deviation ξ_i or the level of slack (Géron, 2012). Because the SVR is fitted on the breaches ξ of the accepted margin of error ε, the model is highly sensitive to scale differences in the input features. To solve this potential source of error we standard scale the data before training the model. To standard scale the data we divide each data point for a feature with the mean of the feature and divide it with the standard deviation of the feature:

$$x_{i-scaled} = \frac{X_i - \bar{x}}{\sigma_x}$$

This way the unit of each of the features will be how many standard deviations away from the mean each data point is.

Because SVR models are so memory efficient, with SVR regression it is possible to convert non-linear relationship to higher-dimension problems to be solved, with the introduction of support kernels. Kernels are a set of mathematical functions that can be added to the algorithm optimization problem in order to project the problem into a high-dimension space. Some of the common kernels are: linear, polynomial, Gaussian/Radial Basis function (RBF) and Sigmoid kernel (Muprhy, 2012). To fit our SVR model we fit an RBF kernel as it is a general-purpose kernel, which is used when there is no clear prior knowledge of the relationships in the data. It is also the most used type of kernel, as it has localized and finite responses across the entire feature axis (Murphy, 2012). The RBF Kernel function is:

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

Where x_j is the projection of x_i into a high dimensional feature space. The RBF kernel introduces a second hyper-parameter in the form of γ. The two hyper-parameters in the SVR-model we fit on the data are therefore C and γ. The higher the value of C the less lenient towards breaches of the margin of error ε the model will be. The higher the value of γ the lower the value of the kernel function will be, which will increase the risk of overfitting, whereas low levels of gamma will make the model more generalized but run the risk of underfitting (Murphy, 2012). To find the optimal values of C and γ we run a grid-search across the two hyper-parameters to find the pair-wise values which yields the highest R^2.

Table 11: Grid search for optimal levels of hyperparameters C and γ

R2 scores SVM	Gamma - values						
C-values	0,1	0,5	1	3	5	10	20
0,1	0,426	0,462	0,487	0,470	0,442	0,410	0,375
0,5	0,445	0,482	0,506	0,489	0,461	0,430	0,394
1	0,462	0,499	0,523	0,506	0,478	0,447	0,411
3	0,477	0,514	0,538	0,521	0,493	0,462	0,426
5	0,487	0,523	**0,547**	0,530	0,530	0,530	0,530
10	0,465	0,501	0,525	0,508	0,480	0,449	0,413
20	0,438	0,501	0,503	0,508	0,480	0,449	0,387

Table 11 show the result of the grid-search for the two variables. The final SVR model we implement will have the hyper-parameter values of C = 5 and $\gamma = 1$,. The final model is trained on the data using the Sci-Kit Learn library module *Support Vector Regression* (Sci-Kit Learn documentation, 2020).

9.6 Ensemble / Voting Regressor

The fourth and last Machine Learning Regression model we set up is a Voting Regressor. Like the Random Forest, a Voting Regressor is an ensemble learning method built on the idea that a collection of several models working together to make a prediction will perform better than a single prediction model alone. But unlike the Random Forest, which is the same model trained on different subsets of the data, a voting regressor is the combination of different regression models trained on the same data. As mentioned, there is no universally better Machine Learning model for prediction. A Voting Regressor is built on the idea that different Machine Learning algorithms make different mistakes (Géron, 2017). Aggregating several predictors into a single predictor has been proven to often improve the outcome, and it is possible to combine several 'weak learners' into a single 'strong learner' (Géron, 2017). This will be true, as long as the models make different types of errors, and one of the models is not a lot more powerful than the others. If the models all make the same kind of error, the benefit of combining them will be limited, as the same mistakes will simply be made by the Voting Regressor. If one of the underlying models is a lot more powerful predictor than the other models, then taking the average of the models will just water out the predicting power of the 'strong learner', and the Voting Regressor will be performing worse than this individual, well-functioning model.

A Voting Regressor is fitted by training each of the different underlying models. When making predictions for each data point, the Voting Regressor will take the predicted value of each of the underlying models and then average it into a single predicted value. When taking the average of

the underlying predictions the main configuration to consider for a Voting Regressor is the weight to give to each underlying model. We train a Voting Regressor using the three above mentioned models: The Ridge Regression, The Random Forest, and the Support Vector Regressor, and assign equal weight to each of the models. The final voting regressor is shown in figure 33. The voting regressor is trained using the Sci-Kit Learn python library module *Voting Regressor*.

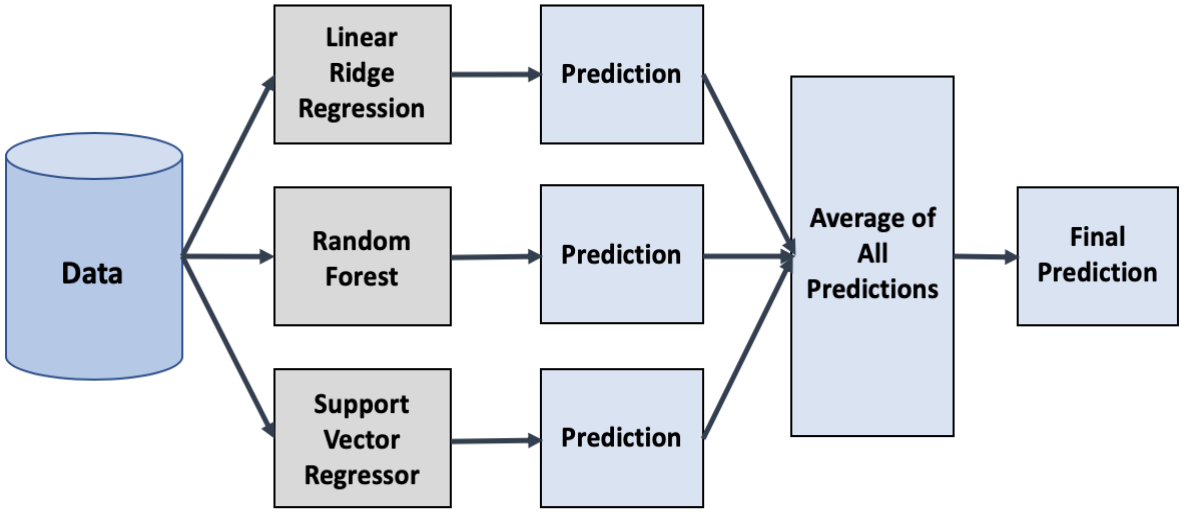


Figure 33: Illustration of the voting regressor setup

9.7 Selecting features - Recursive Feature Elimination

After each model has been set up, we want to derive the final set of features that will yield the best predictive power across the models. Many different methods for optimizing the features used in the final models exists, each with their own benefits and downsides (Géron, 2017). We considered using Principal Component analysis (PCA) to compute a few very important features, as this method has been proved to consistently yield good results (Géron, 2017). However, as PCA analysis will yield features which are in essence a combination of all the underlying features, and the objective of the study is not solely to setup the model that will yield the highest predictive power, but also one that allows for understanding of which features are important for predicting the spreads of high yield bonds, PCA analysis was discarded. Instead we use Recursive Feature Elimination (RFE) with cross-validation to derive at the optimal set of features to use in the final model.

Recursive Feature Elimination is a backward selection of features to use. The selection begins by building a model using the entire set of features and computing the importance of each feature. Table 12 shows the feature importance for the full set of features. The least important features are then dropped from the model and the model is refitted again using the remaining features. This process is then repeated 40 times, since the total number of features is 41. A score for every

combination of features is then computed to find the optimal number (Géron, 2017). The score is calculated using 5-fold cross validation as with the grid-search. This way, the model will never have been tested on the test data before the final test. Figure 34 shows the cross-validation score for each set of features. This is a highly computational requiring approach, therefore we only run the RFE on the Random Forest model, as the process of calculation how important features are for a Random Forest is both less computational requiring and easy to interpret. We then use the set of features from the RFE on the Random Forest model for each of the four Machine Learning models. Using the same features both save computational power and ensure consistency and comparability across the different models. To implement the RFE analysis we use the Sci-Kit learn *RFECV* python library module. As figure 34 shows, 29 features yield the highest cross-validation score. We therefore use the 29 highest ranked features in table 12, for all of the machine learning algorithms.

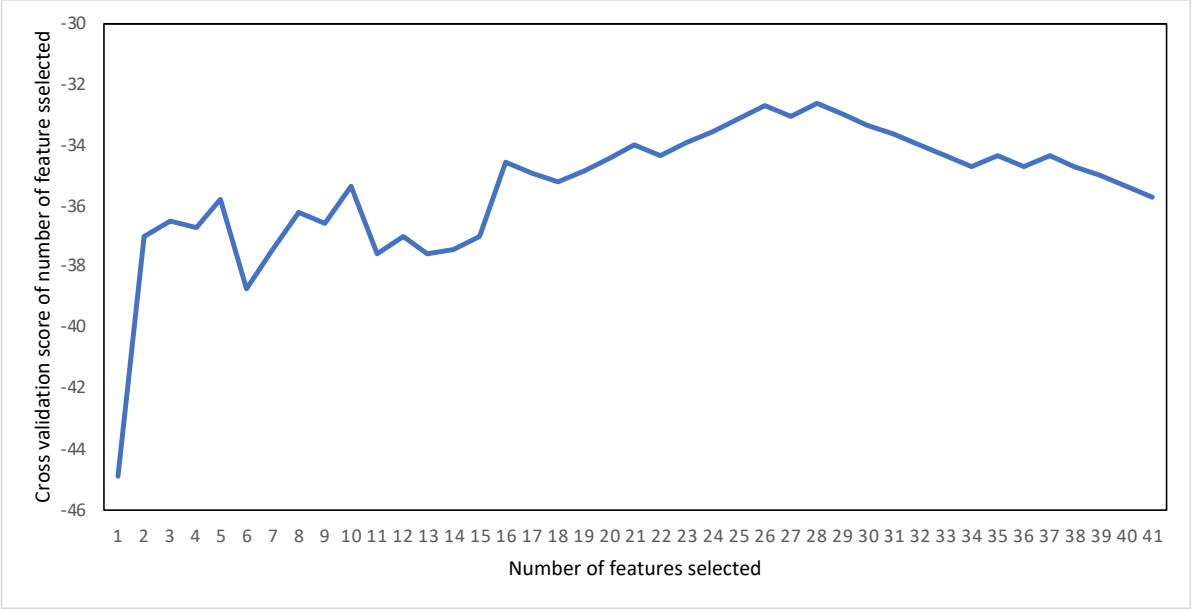


Figure 34: Result of recursive feature selection with cross-validation

Table 12: Feature Importance for model based on Recursive Feature Elimination with Random Forest

Recursive Feature Elimination scoring

Dataset	Public - Issue	Public -31/12/2019	Private
Feature	Ranking	Ranking	Ranking
BB_composite_rating	1	1	n.a.
ICR	1	1	n.a.
industry_Communications	1	2	1
industry_Energy	1	2	1
industry_Industrials	1	1	1
NI_mar_calc	1	1	1
Asset_Tangebility	1	1	n.a.
sentiment_polarity	1	1	1
Sales_3Y_CAGR	1	2	n.a.
adj_net_leverage_calc	1	1	1
topic_2_weight	1	1	1
topic_3_weight	1	1	1
topic_1_weight	1	1	1
geo_South_EU	1	1	n.a.
industry_Consumer Discretionary	1	1	1
industry_Health Care	1	1	2
industry_Technology	1	1	2
WC_to_assets	1	1	n.a.
industry_Utilities	1	1	1
subjectivity	1	1	1
Industry Industrials	2	1	1
Issue_size	2	1	1
ccy_USD	2	2	2
ccy_GBp	2	2	2
adj_EBITDA_mar_calc	2	1	2
geo_Other	2	2	2
geo_US	2	2	2
tot_word	3	2	2
Seniority_SNDB	4	3	n.a.
RoA	5	4	n.a.
cash_ratio	6	5	3
uncer_share	7	6	4
ccy_other	7	7	5
neg_share	8	8	6
Callable_Y	9	9	7
Equity_cushion	10	10	n.a.
lit_share	11	11	9
pos_share	12	12	10
EBITDA_/_Interest_Expense	n.a.	n.a.	1
Operating Cash flow / Net Debt	n.a.	n.a.	1
Topic_4_weight	n.a.	n.a.	1
Topic_5_weight	n.a.	n.a.	1
S&P	n.a.	n.a.	1

9.8 Summarizing the models

The final four models which will be used to predict yield spreads for the test set are the following: a linear Ridge Regressor with a lambda value of 20, a Random Forest with a max depth of 10 and 20 different estimators, A Support Vector Regression model with an RBF kernel and a C-level of 5 and γ – level of 1, and a Voting Regressor combining these three into a single predictor using equal weights to average them.

10. Results – Machine learning regression

The following table presents the result for all machine learning algorithms employed for the private dataset, with spreads at 31-12-2019 as the explanatory variable:

Table 13: Result of final Machine Learning models on EU private dataset

Private dataset

Rating Variable only

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	140,047	294,100	0,213	0,199	142,717	299,705	0,198	0,165
Random Forest	**135,777**	**244,398**	**0,237**	**0,251**	**138,980**	**250,164**	**0,219**	0,184
Support Vector Machine	140,759	295,594	0,209	0,195	144,674	303,816	0,187	0,161
Voting Ensemble	137,378	247,281	0,228	0,242	139,158	292,231	0,218	**0,199**

Rating + quantitative features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	124,388	261,214	0,301	0,501	128,837	270,557	0,276	0,423
Random Forest	**117,092**	**210,765**	**0,342**	**0,603**	121,363	**218,453**	0,318	0,492
Support Vector Machine	124,922	224,859	0,298	0,487	128,303	269,436	0,279	0,411
Voting Ensemble	117,982	247,761	0,337	0,582	**120,295**	252,619	**0,324**	**0,495**

Rating + quantitative features + text features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	81,324	170,780	0,543	0,542	88,798	159,836	0,501	0,508
Random Forest	**71,002**	**127,804**	**0,601**	**0,675**	75,985	159,569	0,573	0,572
Support Vector Machine	75,095	135,172	0,578	0,538	78,832	141,898	0,557	0,543
Voting Ensemble	72,604	152,468	0,592	0,631	**74,384**	**133,890**	**0,582**	**0,592**

In line with expectations, the Random Forest algorithm performs significantly better on the training set than other algorithms. However, it is also evident that this is due to overfitting, as the predictions deteriorate when applied to the test set, even in the case where only rating is used as an independent variable and the model outperforms the other models. Except when the only rating is used as feature, the voting ensemble performs best on the test set. Additionally, adding text features increase explanatory power of the model to a larger degree than adding quantitative

features, in line with results found for the statistical multiple linear regression performed on the full dataset.

The following table presents the result for all machine learning algorithms employed for the public dataset, with spreads at 31-12-2019 as the explanatory variable:

Table 14: Result of final Machine Learning models on EU private dataset at 31/12/2019

EU public at 31/12/2019

Rating Variable only

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	149,123	313,158	0,162	0,174	150,013	315,027	0,157	0,165
Random Forest	**136,488**	**245,679**	**0,233**	**0,244**	144,318	303,068	0,189	0,184
Support Vector Machine	147,877	310,542	0,169	0,167	150,547	316,148	0,154	0,161
Voting Ensemble	140,403	294,847	0,211	0,223	**141,827**	**255,289**	**0,203**	**0,199**

Rating + quantitative features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	150,013	315,027	0,157	0,165	103,390	217,118	0,419	0,423
Random Forest	**144,318**	**303,068**	**0,189**	**0,184**	91,111	164,000	0,488	0,492
Support Vector Machine	150,547	316,148	0,154	0,161	106,415	223,471	0,402	0,411
Voting Ensemble	141,827	255,289	0,203	0,199	**90,221**	**162,398**	**0,493**	**0,495**

Rating + quantitative features + text features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	83,281	149,906	0,532	0,542	93,958	169,125	0,472	0,452
Random Forest	**61,215**	**110,187**	**0,656**	**0,675**	77,765	139,976	0,563	**0,572**
Support Vector Machine	80,612	169,285	0,547	0,538	97,695	205,160	0,451	0,472
Voting Ensemble	66,198	139,015	0,628	0,631	**73,494**	**132,289**	**0,587**	0,571

Similar to the private dataset, the Random Forest algorithm performs better on the training set, while the voting ensemble is the best predictor of spreads for the test set. With R^2 of 0,587 and explained variance of 0,572 the model including textual features explains the largest portion of variance, at very similar levels to that of the private dataset.

The following table presents the result for all machine learning algorithms employed for the public dataset, with spreads at issue as the explanatory variable:

Table 15: Result of final Machine Learning models on EU public dataset, at issue

EU public at issue

Rating Variable only

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	74,206	155,832	0,583	0,569	81,502	171,153	0,542	0,165
Random Forest	72,604	**130,687**	0,592	0,606	79,900	**143,820**	0,551	0,184
Support Vector Machine	76,341	160,316	0,571	0,557	82,035	172,274	0,539	0,161
Voting Ensemble	**67,443**	141,631	**0,621**	**0,607**	**77,053**	161,811	**0,567**	**0,199**

Rating + quantitative features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	65,486	117,875	0,632	0,501	71,536	128,765	0,598	0,423
Random Forest	**57,122**	**102,820**	**0,679**	**0,603**	69,045	144,994	0,612	0,492
Support Vector Machine	71,002	127,804	0,601	0,487	73,494	154,337	0,587	0,411
Voting Ensemble	62,817	131,915	0,647	0,582	**65,486**	**117,875**	**0,632**	**0,495**

Rating + quantitative features + text features

Model setup	Training set				Test set			
Algorithm	MAE	RMSE	R2	Expl Var	MAE	RMSE	R2	Expl Var
Linear Ridge Regression	53,563	96,414	0,699	0,542	63,706	114,672	0,642	0,452
Random Forest	38,615	81,092	0,783	**0,675**	53,029	111,362	0,702	**0,572**
Support Vector Machine	49,114	103,140	0,724	0,538	56,055	117,715	0,685	0,472
Voting Ensemble	**38,082**	**68,547**	**0,786**	0,631	**48,759**	**87,765**	**0,726**	0,571

Unlike the previous datasets, all models manage to achieve R^2 of more than 0,5 using only rating as explanatory variable. The highest score of all datasets is also achieved by the voting regressor using all types of predictors, with R^2 of 0,726.

10.1 Discussion of results

The results clearly demonstrate that textual features of bond prospectuses contain valuable information when predicting bond spreads, with all three datasets achieving the highest level of explained variance when both accounting and textual features are used as explanatory features. The level at which textual features contributes to an increase in prediction power, however, is not equal across the datasets. It appears that more prediction power is gained in the private dataset, which underpins the idea that investors rely on bond prospectuses to a larger degree for private than for public firms, a result that is in accordance with results from the statistical multiple linear regression on the full dataset. As suggested earlier, this could be a result of the large amount of information already available on public firms, meaning investors rely less on the prospectuses to determine credit risk. However, the overall ability to correctly predict spreads is higher for public firms, as the level of explained variance is larger than for the private dataset. This could be due to the fact that information is updated more frequently and in greater detail for publicly traded firms, as some private companies' latest financial statements are dated more than a year earlier than the date the dependent variable is recorded. As with the multiple linear regression, rating has the highest explanatory power for spreads at issue. However, prediction power of the models

improved significantly when adding accounting variables and textual features, which was not the case for the multiple linear regression. There is no definite reason as to why results on this topic differs between model setups, and it is worth noting that the dataset for bond spreads at the time of issuance is both the smallest and least robust of the three datasets. That a prospectus is most relevant at the time of issue seems plausible, but the results of the study cannot reject or confirm this hypothesis conclusively.

Between the different algorithms, several interesting findings are shown in the results. Of the three individual models, the Random Forest outperforms linear Ridge regression and Support Vector regression in almost every instance. This suggest that non-linear relations exist in the data, which would cause linear estimators to fail. Such relationships are well in line with theory, where an example could be cash to assets ratios. Extremely small levels could mean low cash reserves, which would alert investors. On the other hand, very high levels could mean that cash is being used inefficiently. Likewise, leverage levels moving a half a turn, from 2 to 2,5, is much less alarming than a move of the same size at higher levels of leverage, e.g. from 7 to 7,5. However, as is often the case with different algorithms, errors for different datapoints are not equal across models, which explains why the ensemble model outperforms the Random Forest algorithm on the test set in most instances. As such, even models that do not give optimal predictions as a standalone algorithm add prediction power of the ensemble model, increasing the maximum level of accuracy achieved. This tendency is seen commonly seen when using machine learning for text analysis on equity prospectuses, among others in Deaokar & Tao (2015), Deokar et al (2015) and Yan et al (2019). Another interesting finding in line with this is the fact that non-linear algorithms appear to benefit the most from the addition of text features, indicating that relation between bond prospectus content and bond performance is characterized by a non-linear relationship.

From the recursive feature selection, we can determine that the geographical dummy variable for bonds issued in Southern European countries is a good feature for prediction of spreads, confirming the existence of a Southern European premium which was established in the linear regression. Currency dummy variables, however, proved not to be important features when the algorithm predicts spread. Certain industry dummy variables also ranked high in the feature selection. Theoretically, direct effects from industry classification to spreads could relate to its ability to capture the business cycle sensitivity, or beta, of the company. But like textual features, industry variables can also function as a feature capturing underlying risk and information not correctly captured by accounting variables alone. An unexpended finding is the relative importance of net income margin versus EBITDA margin for the Random Forest algorithm. With net leverage being calculated using EBITDA, and EBITDA being accepted as the best measure for a company's ability to meet its interest payments (Fridson, 2018), the finding can appear counterintuitive. While EBITDA margin and net leverage have very direct implications for the perceived security of the bond holder, net income margin signals the health of the company in the

view of equity holders. And since equity holders have decision power over the future development of the firm, low net income margins would likely be followed by changes to the company which brings uncertainty. In this way, net income margin could very well be a good predicter of bond spreads through indirectly derived effects that are not as easily analyzed or quantified by financial analysts as e.g. net leverage. The finding could also simply be a result of the 0,62 correlation between the two measures, meaning the effect of EBITDA margin is captured also by net income margin.

In terms of robustness of results, sensitivity to the training/test split was observed during the study. According to Daumé (2017), this is common for small datasets. As such, replicating the study with more datapoints would be an effective way of minimizing the risk of such sensitivity and increasing the overall robustness of the results. Particularly for yields at issue, the amount of data and the quality of the data resulted in the final dataset being very small. This limits the certainty of any inference drawn from results on the dataset, and as such, allows the hypothesis on bond prospectus relevance across time to be answered less conclusive.

11. Implications for academia and practice

This section will discuss the implications of the results found in this paper for both the literature related to high yield bonds, but also for practitioners, including an assessment of how the analytical framework presented in the study could be useful for Capital Four. The discussion will be split in three parts: A brief review of the full analytical pipeline including how it could be implemented in practice, a section on implications for academia, and a section on implication for practitioners.

11.1 Overview of the analytical pipeline

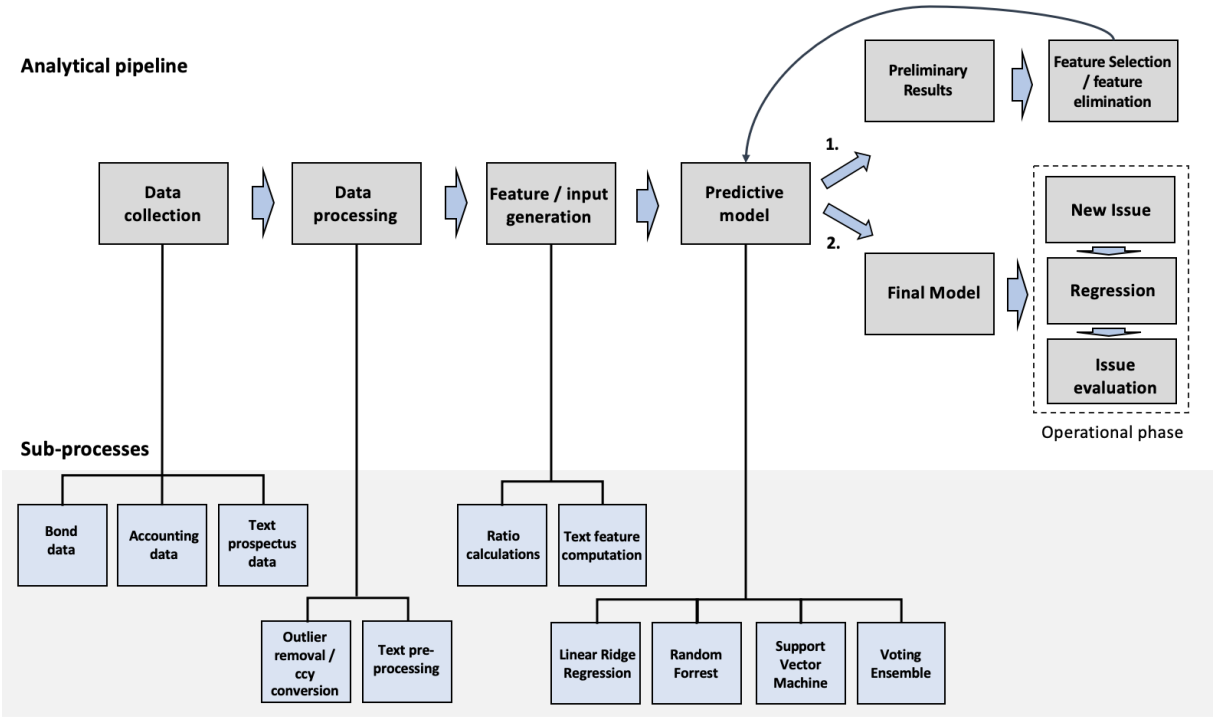


Figure 35: Overview of the complete analytical pipeline

Figure 35 shows the full analytical process of the final model for predicting high yield bond spreads. First, all the data is collected. Here the bond-specific information regarding the individual bonds are collected from Bloomberg, accounting data for the issuing firm is collected from 9Fin and Bloomberg and the prospectuses are collected from Bloomberg and 9Fin as well. Secondly the data is being processed so it is ready for analysis. For the quantitative data, we convert everything to the same currency and scale, outliers are removed etc. For the text data the text is processed through the text pre-processing steps described in the text-processing section. Afterwards the features based on the data is calculated. For the quantitative data, the financial ratios used are calculated and categorical variables are turned into individual dummy variables. For the prospectuses, each of the three text features: Fundamental word score, sentiment analysis and topic modelling is computed for each text. With all the data processed and the features calculated the final predictive models are setup.

We set up three different machine learning models and a Voting Ensemble which combines each of these individual models. We then compute the preliminary scores of these models, and fine tune them using grid-search and recursive feature elimination with 5-fold cross-validation to derive at the final set of features and the setting of hyper-parameters for each of these models. When the model settings and the set of features has been fine-tuned, the final model is then trained on the test data and each of the four machine learning models are evaluated.

The next step, should the framework be adopted by practitioners such as Capital Four, would be to include the best of the final models in an *operational phase*. Here, the model can be used as a screening tool or as part of the investment decision. In the operational phase, when a new issue comes into the market, the data on the new issue would be collected and processed. The features will be computed and inputted into the final model, which will then return a prediction of the yield spread, which can be compared to the spread being offered at issue. This comparison can be used in the investment decision process or as a screening tool, flagging issues where the spread differs greatly from prediction. Finally, each new prediction and result in the operational phase should be saved and added to the dataset, as to further improve the performance of the model.

11.2 Implications for academia

11.2.1 Contributions of the model

A main contribution of this model to the overall literature on high yield bond spreads is the development of a full analytical pipeline and model framework that can be used for high yield bond spread prediction. To the knowledge of the authors, it is the first applications of a full machine learning model set up within the academic realm of high yield bonds in Europe. The framework thus sheds light on the potential to gain further information by the adoption of more advanced, computationally heavy models that do not rely on simple linear relationships. The machine learning models can predict high yield bond spreads through a much larger variety of feature inputs and from more complex features than the previous studies mentioned in the literature review. The full analytical pipeline provides a model that can easily be adopted by future researchers who want to examine the causes of high yield bond spreads and want to apply a machine learning framework to do so.

This paper also contributes to the literature by building a model that allows to examine high yield bonds through textual features. Through the model, the content of the high yield bond prospectuses now figures as features that are used to explain and predict the risk spread. It thereby takes new practices, which have already been tested in the academic realm of equities and apply them to the academic field of corporate bond spreads. As we calculate three different types of features reflecting the underlying text of the corporate bond prospectuses, we provide a lens to better understand which part of the bond prospectus texts are relevant for yield pricing. As shown by the recursive feature selection as well as the multiple linear regression, topic weights and sentiment scores proved to be stronger predictors than fundamental word scores.

Lastly, the analytical pipeline, developed to predict risk spreads of high yield bonds through a set of collected data, can be generalized to use in domains other than high yield bonds. The framework could easily work for studies of IPOs pricing, price-movements following financial reports such as annual or quarterly reports, how news affect financial markets and similar studies that seeks to combine quantitative data and textual data to explain pricing in the financial markets. A key

challenge still remaining to effectively generalize the full analytical pipeline, is effectively obtaining labeled data to train the predictive models on.

11.2.2 Assumptions and limitations of the model

One of the first limitations of the model, as with most statistical models is that it only shows correlation and not causation. Even though we have found certain variables to have high explanatory power in the prediction of the yield spread, we cannot be sure that those variables are causing the spread. The model is set up with a finite number of features, to limit the degrees of freedom lost and to not have the dimensions of the dataset explode. As a result, the strong explanatory power of some variables may simply stem from a high correlation with a third variable, that is not included in the model, which have the actual causal effect on the spreads. The model still can provide important insight as to which textual, bond, and accounting features financial analyst as well as the managers of issuing companies should pay attention to, but it cannot confirm any causal relationships between them and high yield bond spreads.

Another assumption of the model set up in this paper is that current and past data works as a predictor of the future. High yield bond spreads are assumed to be reflecting the investors' expectations of the future of this bond and the market as a whole. The whole model is therefore built on the assumptions that statements given in the prospectuses and historical accounting data will hold some predictive power of future events. This has been proven historically, as e.g. accounting data often falls within a certain range of previous years accounting data. But this assumption may only hold in normal and general circumstances, as the current Covid-19 crisis has proven. The crisis has caused situation where many historical indicators, which have normally been strong predictors of future performance, have proven to be completely irrelevant predictors.

A model such as the one developed for this paper is also only as good as the data it is built on. We had to limit the number of features and feature complexity because of the limited size of the final dataset, as we did not want to lose to many degrees of freedom. A larger dataset would allow for more complex feature set, e.g. using brute force machine learning on raw TF-IDF matrix inputs with several thousand features, to discover hidden features in the texts. Furthermore, the model is built under the assumption that the data it is trained on is accurate. It assumes that the spreads obtained from the Bloomberg platform are the spreads the bonds would trade at if an investor tried to invest in them. If they are not, the predictions made by the model will no longer be accurate. The same goes for the historical accounting and bond data. To make sure the findings in this paper is generalizable, one could re-run the model setup on data obtained from another source i.e. FactSet or another similar trading platform.

11.2.3 Suggestion for further research

The findings of this paper combined with the assumptions and limitations discussed above raise several interesting next steps for further research. First it would be interesting to run the same analytical pipeline on a larger European high yield bond set and compare the results to make sure

the results are also consistent on very large data set. Another potential further study would be to apply the analytical framework to the study of US high yield bonds. The reporting requirements for the prospectuses i.e. the language used in them as stated in the 'Plain English Rule' enforced by the SEC, are different in the US (Bartov, 2011). It would therefore be an interesting comparison to compare the results of a study on US high yield bonds with the findings of this paper.

Furthermore, this paper only investigated spreads, which are investors' expectations of the future. This is not a measure of market outperformance and consequently in line with the Efficient Market Hypthesis (Fama, 1970). However, several structural reasons may cause the Efficient Market Hypothesis to not hold in the European high yield bond market. The liquidity is low, trades are only done in large sizes and over the phone, it takes a lot of time for news to settle in, and market algorithms have yet to take over the investment decision process. It would therefore be an interesting next step to see whether the framework developed in this paper could be used to predict some measure of market outperformance for high yield bonds. Such a research project would have to setup a structure to run 'alpha' calculations on high yield bonds in addition to the framework developed in this paper. The alphas calculations could then be adapted to the framework either directly as another regression problem, or it could be converted into a classification problem, with positive alphas as one class and negative alphas as another class. Setting it up as a classification problem would enable a larger array of machine learning models to be applied.

Lastly, this paper only tries to predict and explain the regular credit risk found in high yield bonds throughout most of the financial cycle. However, as depicted on figure 36 the normal differences and changes in the high yield risk spread is dwarfed by the yield in extreme situations such as the Covid-19 crisis. A further next step of research could be to use the model developed in this paper to try to explain the spread movement of individual high yield bonds in extreme crisis such as the 2001 Dot com bubble, the 2008 financial crisis or the 2020 Covid-19 crisis. This would be interesting for high yield bonds, as many of the loses found in high yield bonds in newer times are caused by such events. Figure two shows the default rate of European high yield bonds, which shows that defaults are clustered around the two last financial crises, and indicators predict that this crisis might hit even harder (S&P Global, 2019)

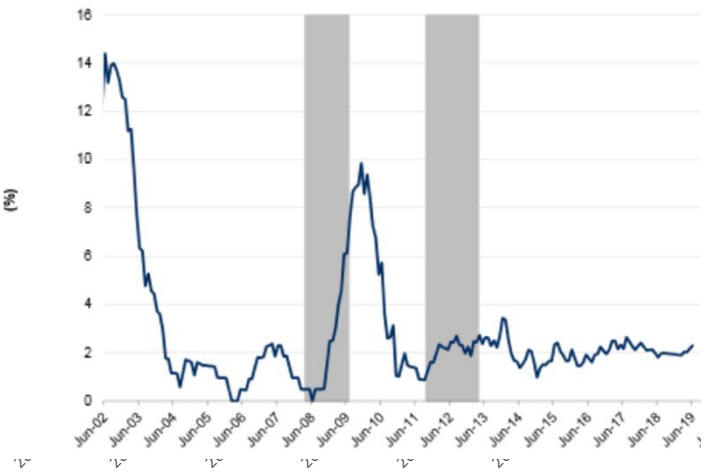


Figure 36: Historic credit spreads of High Yield bonds

The purpose of this paper was not purely to create an academic model with academic implications, but also to create a model which can be used in practice and create value for high yield asset managers such as Capital Four Management. The first clear implication for practitioners of the findings in this paper is that the prospectuses hold information relevant to the pricing of bonds. They should therefore be included in the investment decision process. Furthermore, if the model developed in this paper was perfectly setup with every relevant feature included, the investment process could be automated. This is seen in the world of equities where most of the trading and investment process is now undertaken by algorithms (Frasincar et al, 2013). However, the high yield world is constrained by low volumes of data, which was also the biggest constraint of this paper. The result of the model trained still had significant RMSE and MAE with values in the very best testing results of 0,877% and 0,487% which is still significant margins of error when choosing whether or not to invest in high yield bonds.

The models do consequently not have the predictive power to automate the full investment decision-making process. However, it is the believe of the authors that the framework is still useful in several ways for asset managers such as Capital Four Management. Firstly, it can be used in the screening stage of the investment process. If a bond is predicted to trade at a spread significantly lower than the spread offered in the market, it can be flagged as worthwhile looking into. The same goes for bonds trading at much lower spreads as predicted, as these can quickly be discarded as not worth spending analyst resources on. Secondly, the model can function as a validation of the investment decision process. In the world of asset management, portfolio managers make decisions ex-ante which will not always turn out to be great decisions ex-post. They often face the task of proving that the investment decision was still a good idea ex-ante, given the information known at that time. The model can be used as another analytical tool and documentation of the ex-ante decision-making process.

Lastly, the model can be used as a form of stored memory. If an asset manager would label each investment decision and store it in a proper database with the prospectuses that have been analyzed by the company's team of analysts. A model such as this could then be trained gradually more powerful over time and serve as a collective memory across the analyst team, as each feature along with the investment verdict label would be stored properly. In the case of Capital Four Management, a potential setup would be to keep a database containing the underlying features of each investment case analyzed, as well as the decision made at each step of the investment process, from initial screening to investment review.

To increase the benefits from the model developed in this paper in practice, the biggest challenge that must be faced is to automate the data collection process. For this study, the task of gathering bond prospectuses was a very manual and time-consuming task. As such, to really benefit of automated feature generation from text, the collection of the text used as input would need to be automated as well.

12. Conclusion

This paper has investigated whether information content contained in the prospectuses of European high yield bonds can explain yield spreads. High yield bonds carry significant credit risk, and a deep understanding of the underlying company is therefore needed to correctly price securities. To provide information about the company and the security, large amounts of text are written in bond prospectuses. While it is common practice in the world of equity investing to use prospectus text in a structured manner for modelling, the niche nature of the high yield market means that such analysis has not been implemented yet.

After extracting this text for two datasets containing European high yield bonds issued by public and private companies between 2014 and 2019, we performed unsupervised clustering of the topics found using LDA topic modelling and used topic weights for each prospectus as model input. We also created a fundamental word score specifically developed for financial language, and lastly, we performed sentiment analysis, using ratings for sentiment and subjectivity as model input. Combined with accounting ratios commonly accepted to be determinants of bond spread, as well as standard bond characteristics variables, these constitutes the explanatory variables for the study.

A multiple linear regression was performed, first with only rating as the explanatory variable, then adding quantitative variables, and lastly adding textual features from bond prospectuses. Similarly, we set up four machine learning algorithms to predict spreads on an unseen test dataset. The four algorithms used were a linear Ridge regressor, a Random Forest algorithm, a Support Vector regressor and a voting classifier consisting of the previous three model, weighted equally.

We found a significant correlation between topic weights and bond spread, with textual features adding explanatory power to a multiple linear regression model in addition to common quantitative metrics. The information gain is largest for private companies, indicating an increased importance of text in bond prospectuses for firms that do not regularly publish large amounts of information. Additionally, the results show that rating is the most significant determinant of spread at the time of issue, with quantitative and qualitative features not adding explanatory power.

For the machine learning models, the voting classifier performed best in almost every instance. Additional prediction power was added with the inclusion of text features from bond prospectuses for both the private and public dataset, both when predicting spreads at 31-12-2019 and when predicting spreads at the time of issuance. Of the individual models, the Random Forest algorithm performed best, which indicates that the relationship between bond prospectus features and yield spread is of a complex, non-linear nature. In line with results from the multiple linear regression, topic weights were a significant determiner of spreads, while fundamental word score was a less powerful predictor.

The findings contribute to academia by extending the relatively well-developed analysis of equity prospectuses for equity pricing to the universe of high yield investment, finding that text data can also be a valuable model input when modelling bond spreads. Additionally, the study presents a framework for industry practitioners, which could increase accuracy of investment decisions and help structure the documentation of the investment process, by quantifying inherently qualitative information.

13. References

9Fin. (2020). *Database of High Yield Companies*. https://9fin.com/about

Abdullah, S. S., Rahaman, M. S., & Rahman, M. S. (2013). Analysis of stock market using text mining and natural language processing. *2013 International Conference on Informatics, Electronics and Vision, ICIEV 2013*, *May*. https://doi.org/10.1109/ICIEV.2013.6572673

AFME (Association for Financial Markets in Europe). (2020). *European High Yield & Leveraged Loan Report*. https://www.afme.eu/Portals/0/DispatchFeaturedImages/-2High Yield and Leveraged Loan Report Q4 2019-2-3.pdf

Alpha Wise. (2020). *High Yield Market Statistics*. https://www.wisealpha.com/statistics#market-size

Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Amato, J. D., & Remolona, E. M. (2012). The Credit Spread Puzzle. *SSRN Electronic Journal*, *December*, 51–64. https://doi.org/10.2139/ssrn.1968448

Anderson, R., & Sundaresan, S. (2000). A comparative study of structural models of corporate bond yields: An exploratory investigation. *Journal of Banking and Finance*, *24*(1–2), 255–269. https://doi.org/10.1016/S0378-4266(99)00059-X

Ashby, D., & Kumar, N. (1996). *A Comparison of Neural Networks and Classical Discriminant Analysis in Anticipating Default among High-Yield Bonds*.

Ashby, D., & Kumar, N. (1996). *A Comparison of Neural Networks and Classical Discriminant Analysis in Anticipating Default among High-Yield Bonds*.

Askin, R. G., Cresswell, S. H., Goldbreg, J. B. G., & Vakharia, A. j. (1991). ContentServer (1).pdf. In *Physical Therapy* (Vol. 29, pp. 1081–1100). https://doi.org/10.1177/017084068800900203

Balakrishnan, K., & Bartov, E. (2011). Analysts ' Use of Qualitative Earnings Information : Evidence from the IPO Prospectus ' s Risk Factors Section Analysts ' Use of Qualitative Earnings Information : Evidence from IPO Prospectus ' s Risk Factors Section. *October*, *February*.

BAML. (2016). *The ice Indiecy*. https://www.theice.com/publicdocs/BoA_Change_Reference_Data_Source.pdf

Bao, J., Pan, J. U. N., & Wang, J. (2016). *The Illiquidity of Corporate Bonds Published by : Wiley for the American Finance Association The Illiquidity of Corporate Bonds*. *66*(3), 911–946.

Berk, J., & DeMarzo, P. (2016). *Corporate Finance* (4. ed.).

Bertocini, S., Feltus, A., Monaghan, K. J., & Fiorot, L. (2020). *No Title*.

Bhardwaj, A., Narayan, Y., Vanraj, Pawan, & Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *Procedia Computer Science*, *70*, 85–91. https://doi.org/10.1016/j.procs.2015.10.043

Bird, S., Klein, E., & Loper, E. (2009). *No Title* (J. Steele (ed.); 1st ed.). O'Reilly.

Black, F., & Cox, J. C. (1976). Valuing Corporate Securities: Some Effects of Bond Indenture Provisions. *The Journal of Finance*, *31*(2), 351–367. https://doi.org/10.1111/j.1540-6261.1976.tb01891.x

Blochwitz, S., & Nyberg, M. (2000). Benchmarking Deutsche Bundesbank ' s Default Risk Model , the KMV □ □ Private Firm Model □ □ and Common Financial Ratios for German Corporations Abstract : *Portfolio The Magazine Of The Fine Arts*.

Blume, M. E., Lim, F., & Mackinlay, A. C. (1998). The declining credit quality of U.S. corporate debt: Myth or reality? *Journal of Finance*, *53*(4), 1389–1413. https://doi.org/10.1111/0022-1082.00057

Bodie, Z., Kane, A., & Marcus, A. J. (2018). *Investments* (11. ed.).

Brealy, R. A., Myers, S. C., & Marcus, A. J. (2015). *No Title* (eigth). McGraw-Hill Education.

Bruner, R. F. (1988). The Use of Excess Cash and Debt Capacity as a Motive for Merger. *The Journal of Financial and Quantitative Analysis*, *23*(2), 199. https://doi.org/10.2307/2330881

Butera, G., & Faff, R. (2006). An integrated multi-model credit rating system for private firms. *Review of Quantitative Finance and Accounting*, *27*(3), 311–340. https://doi.org/10.1007/s11156-006-9434-7

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, *63*(6), 2899–2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x

Campbell, J. Y., & Taksler, G. B. (2003). Equity Volatility and Corporate Bond Yields. *The Journal of Finance*, *58*(6), 2321–2350. https://doi.org/10.1046/j.1540-6261.2003.00607.x

Campbell, J. C., Hindle, A., & Stroulia, E. (2015). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. *The Art and Science of Analyzing Software Data*, *3*, 139–159. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Capital Four Management. (2020). *Capital Four Mangement Awards*. https://capital-four.com/awards/

Castagnolo, F., & Ferro, G. (2014). Models for predicting default: towards efficient forecasts. *Journal of Risk Finance*, *15*(1), 52–70. https://doi.org/10.1108/JRF-08-2013-0057

Chen, L., Lesmond, D. A., Wei, J., Chen, L., Lesmond, D. A., & Wei, J. (2016). *Corporate Yield Spreads and Bond Liquidity Published by : Wiley for the American Finance Association Stable URL : http://www.jstor.org/stable/4123458 Corporate Yield Spreads and Bond Liquidity*. *62*(1), 119–149.

Cherubini, U., & Lunga, G. Della. (2001). Liquidity and credit risk. *Applied Mathematical Finance*, *8*(2), 79–95. https://doi.org/10.1080/13504860110061013

Collin-Dufresne, P., & Goldstein, R. S. (2001). Do Credit Spreads Reflect Stationary Leverage Ratios? *The Journal of Finance*, *56*(5), 1929–1957. https://doi.org/10.1111/0022-1082.00395

Collin-Dufresne, P., & Goldstein, R. S. (2001). Do credit spreads reflect stationary leverage ratios? *Journal of Finance*, *56*(5), 1929–1957. https://doi.org/10.1111/0022-1082.00395

Daumé, H. (2017). *A course in machine learning*. https://doi.org/10.1007/SpringerReference_35834

Deumes, R. (2008). Corporate risk reporting: A content analysis of narrative risk disclosures in prospectuses. *Journal of Business Communication*, *45*(2), 120–157. https://doi.org/10.1177/0021943607313992

Dick-Nielsen, J., Feldhütter, P., & Lando, D. (2012). Corporate bond liquidity before and after the onset of the subprime crisis. *Journal of Financial Economics*, *103*(3), 471–492. https://doi.org/10.1016/j.jfineco.2011.10.009

Elton, E. J., Gruber, M. J., Agrawal, D., & Mann, C. (2010). Explaining the rate spread on corporate bonds. *Investments and Portfolio Performance*, *LVI*(1), 21–51. https://doi.org/10.1142/9789814335409_0003

Ericsson, J., Jacobs, K., & Oviedo, R. (2009). The Determinants of Credit Default Swap Premia Author (s): Jan Ericsson , Kris Jacobs and Rodolfo Oviedo Published by : Cambridge University Press on behalf of the University of Washington School of Business Administration Stable URL : http://www.jstor. *Journal of Financial and Quantitative Analysis*, *44*(1), 109–132. https://doi.org/: 1 0. 1 0 1 7/S0022 1 0900909006 1

Fabozzi, F. J. (2013). *Bond Markets, Analysis and Strategies Global Edition* (8. ed.).

Falkenstein, E. G., Boral, A., & Carty, L. V. (2005). RiskCalc for Private Companies: Moody's Default Model. *SSRN Electronic Journal*, *May*. https://doi.org/10.2139/ssrn.236011

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2011). The Incremental Information Content of Tone Change in Management Discussion and Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1126962

Fishe, R. P. H., North, D., & Smith, A. (2015). *Prospectus Words , Parts of Speech , and Uncertainty in IPO Pricing Prospectus Words , Parts of Speech , and Uncertainty in IPO Pricing*.

FRED, S. L. F. (2020). *ICE BofA Euro High Yield Index Effective Yield (BAMLHE00EHYIEY)*. https://fred.stlouisfed.org/series/BAMLHE00EHYIEY

Fridson, M. (n.d.). *High-Yield Analysis*.

Fridson, M. S., & Cherry, M. A. (2020). *Tchnical Notes*. *46*(4), 61–67.

Fridson, M. S., & Gao, Y. (1996). Primary versus secondary pricing of high-yield bonds. *Financial Analysts Journal*, *52*(3), 20–27. https://doi.org/10.2469/faj.v52.n3.1992

Fridson, M. S., & Garman, M. C. (1998). Determinants of spreads on new high-yield bonds. *Financial Analysts Journal*, *54*(2), 28–39. https://doi.org/10.2469/faj.v54.n2.2163

Fridson, M., Yang, Y., & Wang, J. (2016). Seniority Differentials in High Yield Bonds: Evolution, Valuation, and Ratings. *Journal of Applied Corporate Finance*, *28*(4), 68–72. https://doi.org/10.1111/jacf.12207

Fung, G. P. C., Yu, J. X., & Lam, W. (2002). *News Sensitive Stock Trend Prediction* (pp. 481–493). https://doi.org/10.1007/3-540-47887-6_48

Gaillard, N. (2012). Fitch, Moody's, and S&P Sovereign Ratings and EMBI Global Spreads: Lessons from 1993–2007. In *A Century of Sovereign Ratings* (Issue April 2009, pp. 149–170). Springer New York. https://doi.org/10.1007/978-1-4614-0523-8_9

Géron, A. (2017). *Hands On Machine Learning*. O'Reilly Media. https://doi.org/10.3389/fninf.2014.00014

Gottfried, M. (2018, June 13). No Title. *WSJ*.

Grøstad, K. N., & Mjøs, A. (2013). *Predicting default in the Norwegian High Yield bond market A study of defaults in the years 2006-2013 Master thesis in Financial Economics*.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques* (third). https://doi.org/10.1109/ICMIRA.2013.45

Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *Review of Financial Studies*, *23*(7), 2821–2864. https://doi.org/10.1093/rfs/hhq024

Huang, J.-Z., & Huang, M. (2011). How Much of Corporate-Treasury Yield Spread Is Due to Credit Risk?: A New Calibration Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.307360

Huang, J.-Z., & Huang, M. (2012). How Much of the Corporate-Treasury Yield Spread Is Due to Credit Risk? *Review of Asset Pricing Studies*, *2*(2), 153–202. https://doi.org/10.1093/rapstu/ras011

Huffman, S. P., & Ward, D. J. (1996). The prediction of default for high yield bond issues. *Review of Financial Economics*, *5*(1), 75–89. https://doi.org/10.1016/S1058-3300(96)90007-5

I., A. (2000). Predicting financial distress of companies: revisiting the Z-Score and ZETA® models. In *Handbook of Research Methods and Applications in Empirical Finance*. Edward Elgar Publishing. https://doi.org/10.4337/9780857936097

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing Derivatives on Financial Securities Subject to Credit Risk. *The Journal of Finance*, *50*(1), 53–85. https://doi.org/10.1111/j.1540-6261.1995.tb05167.x

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*(3), 712–729. https://doi.org/10.1016/j.jfineco.2013.08.018

John, K., Ravid, S. A., & Reisel, N. (2010). The Notching Rule for Subordinated Debt and the Information Content of Debt Rating. *Financial Management*, *39*(2), 489–513. https://doi.org/10.1111/j.1755-053X.2010.01081.x

Khurana, I. K., & Raman, K. K. (2003). Are Fundamentals Priced in the Bond Market? *Contemporary Accounting Research*, *20*(3), 465–494. https://doi.org/10.1506/MTEM-T25T-BCJX-57NC

Kim, I. J., Ramaswamy, K., & Sundaresan, S. (1993). Does Default Risk in Coupons Affect the Valuation of Corporate Bonds?: A Contingent Claims Model. *Financial Management*, *22*(3), 117. https://doi.org/10.2307/3665932

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management*, *12*(1), 29–41. https://doi.org/10.1002/isaf.239

Kovner, A., & Wei, C. (Jason). (2012). The Private Premium in Public Bonds. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2018441

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., & Steinbrecher, M. (2016). *Computational Intelligence: A Methodological Introduction*. https://doi.org/10.1007/978-1-4471-5013-8

Kulshrestha, R. (2019). A Beginner's Guide to Latent Dirichlet Allocation(LDA). *Towards Data Science*. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Laeven, L., & Levine, R. (2008). Complex ownership structures and corporate valuations. *Review of Financial Studies*, *21*(2), 579–604. https://doi.org/10.1093/rfs/hhm068

Lando, D. (1998). On Cox processes and risky bonds. *Review of Derivatives Research*, *2*, 99–120.

Learn, S.-K. (2020). *Sci-Kit Learn Documentation*. https://scikit-learn.org/stable/

Lesmond, D. A., Ogden, J. P., & Trzcinka, C. A. (2017). *The Society for Financial Studies A New Estimate of Transaction Costs Author (s): David A . Lesmond , Joseph P . Ogden and Charles A . Trzcinka Source : The Review of Financial Studies , Vol . 12 , No . 5 (Winter , 1999), pp . 1113-1141 Published by : 12*(5), 1113–1141. https://doi.org/10.1093/rfs/hhrl33

Li, H., McCarthy, J., & Pantalone, C. (2014). High-yield versus investment-grade bonds: less risk and greater returns? *Applied Financial Economics*, *24*(20), 1303–1312. https://doi.org/10.1080/09603107.2014.925049

Lie, C. M., & Nielsen, M. L. (2015). *Empirical Studies of Credit Spreadsin the European High Yield Market*. Copenhagen Business School.

Lin, H., Wang, J., & Wu, C. (2011). Liquidity risk and expected corporate bond returns. *Journal of Financial Economics*, *99*(3), 628–650. https://doi.org/10.1016/j.jfineco.2010.10.004

Longstaff, F. A., Mithal, S., & Neis, E. (2005). Corporate yield spreads: Default risk or liquidity? New evidence from the credit default swap market. *Journal of Finance*, *60*(5), 2213–2253. https://doi.org/10.1111/j.1540-6261.2005.00797.x

LONGSTAFF, F. A., & SCHWARTZ, E. S. (1995). A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. *The Journal of Finance*, *50*(3), 789–819. https://doi.org/10.1111/j.1540-6261.1995.tb04037.x

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307–326. https://doi.org/10.1016/j.jfineco.2013.02.017

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Lynch, B. of A. M. (2015). *The Ice Indicy*. https://www.theice.com/publicdocs/BoA_Change_Reference_Data_Source.pdf

Maciel, L., Gomide, F., & Ballini, R. (2016). A differential evolution algorithm for yield curve estimation. *Mathematics and Computers in Simulation*, *129*, 10–30. https://doi.org/10.1016/j.matcom.2016.04.004

Mahajan, A., Dey, L., & Haque, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market. *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 423–426. https://doi.org/10.1109/WIIAT.2008.309

McGee, L. W., & Spiro, R. L. (1988). ContentServer (2).pdf. In *The Marketing Concept in Perspective* (Vol. 13, Issue 22, pp. 5416–5421). https://doi.org/10.1257/0002828041464551

Merton, R. C. (1974). *American Finance Association On the Pricing of Corporate Debt : The Risk Structure of Interest Rates Author (s): Robert C . Merton Source : The Journal of Finance , Vol . 29 , No . 2 , Papers and Proceedings of the Thirty- Second Annual Meeting of the A*. *29*(2).

Mincer, J. (1958). The University of Chicago Press http://www.jstor.org/stable/10.1086/667722 . *American Journal of Sociology*, *66*(3), 281–302.

Mittermayer, M., & Knolmayer, G. (2006). NewsCATS: A News Categorization and Trading System. *Sixth International Conference on Data Mining (ICDM'06)*, 1002–1007. https://doi.org/10.1109/ICDM.2006.115

Munjal, P., Narula, M., Kumar, S., & Banati, H. (2018). Twitter sentiments based suggestive framework to predict trends. *Journal of Statistics and Management Systems*, *21*(4), 685–693. https://doi.org/10.1080/09720510.2018.1475079

Murphy, K. (2012). Machine Learning - A Probabilistic Perspective. In *Chance Encounters: Probability in Education*. https://doi.org/10.1007/978-94-011-3532-0_2

Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. *2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010*, *4*(March), 256–260. https://doi.org/10.1109/ICCAE.2010.5451705

Norvig, P., & Russell, S. (2010). *Artificial intelligence—a modern approach*. https://doi.org/10.1017/S0269888900007724

Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2014). An Automated Framework for Incorporating News into Stock Trading Strategies. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 823–835. https://doi.org/10.1109/TKDE.2013.133

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADMIRAL: A Data Mining Based Financial Trading System. *2007 IEEE Symposium on Computational Intelligence and Data Mining*, 720–725. https://doi.org/10.1109/CIDM.2007.368947

Radim Řehůřek. (2020). *Gemsim Documentation*. https://radimrehurek.com/gensim/

Reilly, F. K., Wright, D. J., & Gentry, J. A. (2010). An analysis of credit risk spreads for high yield bonds. *Review of Quantitative Finance and Accounting*, *35*(2), 179–205. https://doi.org/10.1007/s11156-009-0162-7

S&P. (2019). *European high yield default rate to rise to 2.8% by June 2020 (S&P)*. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/leveraged-loan-news/european-high-yield-default-rate-to-rise-to-2-8-by-june-2020

S&P Global. (2020). *High Yield Bond Primer*. https://www.spglobal.com/marketintelligence/en/pages/toc-primer/hyd-primer#sec1

S&P Global. (2017). *2017 Annual European Corporate Default Study And Rating Transitions*. https://www.allnews.ch/sites/default/files/files/20180905_SP_2017-Annual-European-Corporate-Default-Study.pdf

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, *27*(2), 1–19. https://doi.org/10.1145/1462198.1462204

Science, C. (2014). *Twitter Data Predicting Stock Price Using Data Mining Techniques*. *4*(6), 9622.

Shekhawat, B. S. (n.d.). *Sentiment Classification of Current Public Opinion on BREXIT : Naïve Bayes Classifier Model vs Python ' s TextBlob Approach Bhupender Singh Shekhawat National College of Ireland Supervisor :*

Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics* (third). Person Education limited.

Stone Water. (2015). *The Globalization of the High Yield Market*. https://www.shiplp.com/wp-content/uploads/Globalization-of-the-HY-Market-March-2015.pdf

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, *48*, 272–281. https://doi.org/10.1016/j.irfa.2016.10.009

Tao, J., & Deokar, A. V. (2015). Text mining for studying management's confidence in IPO prospectuses and IPO valuations. *2015 Americas Conference on Information Systems, AMCIS 2015*, 1–11.

Tao, J., Deokar, A. V., & Deshmukh, A. (2018). Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. *Journal of Business Analytics*, *1*(1), 54–70. https://doi.org/10.1080/2573234x.2018.1507604

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance*, *63*(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

TextBlob. (2020). *TextBlob: Simplified Text Processing*. https://textblob.readthedocs.io/en/dev/

Trainor, W. J. (2010). Performance measurement of high yield bond mutual funds. *Management Research Review*, *33*(6), 609–616. https://doi.org/10.1108/01409171011050217

Tufts, C. (2019). *The Little Book of LDA*. https://ldabook.com/

VanderPlas, J. (2016). *Python Data Science Handbook* (D. Schanafeldt (ed.); First). O'Reilly Media.

Wang, Y., Dwyer, D., & Zhao, J. Y. (2014). *RiskCalc Banks v4.0 Model*. *July*, 1–8. http://www.moodysanalytics.com/~/media/Insight/Quantitative-Research/Default-and-Recovery/2014/2014-01-07-RiskCalc-US-Banks-v4-Model.ashx

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Yan, Y., Xiong, X., Meng, J. G., & Zou, G. (2019). Uncertainty and IPO initial returns: Evidence from the Tone Analysis of China's IPO Prospectuses. *Pacific Basin Finance Journal*, *57*(April 2018), 101075. https://doi.org/10.1016/j.pacfin.2018.10.004

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89–97. https://doi.org/10.1016/j.knosys.2013.01.001

Zhai, Y., Hsu, A., & Halgamuge, S. K. (n.d.). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. In *Advances in Neural Networks – ISNN 2007* (pp. 1087–1096). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72395-0_132

Zhang, T.-W., & Wu, W.-H. (2014). The asymmetric predictability of high-yield bonds. *The North American Journal of Economics and Finance*, *29*, 146–155. https://doi.org/10.1016/j.najef.2014.06.001

14. List of Tables

15. List of Figures