Stock and Corporate Bond Selection using Text

by

Robbert-Jan 't Hoen (400200)

A thesis submitted in partial fulfillment for the degree of
Master of Science in Econometrics and Management Science
Quantitative Finance specialization

Erasmus School of Economics
Erasmus University Rotterdam

Supervised by:
Dr. M. van de Velden (Erasmus University Rotterdam)
Dr. P. Houweling (Robeco)
F. Muskens (Robeco)
Dr. J. van Zundert (Robeco)

Co-reader:
Dr. F. Frasincar (Erasmus University Rotterdam)

September 11, 2019

Abstract

This thesis finds that the text in annual and quarterly reports of U.S. listed companies is related to the cross-section of average stock and corporate bond price volatility and returns even after controlling for other known risk factors. This effect can be attributed to underreaction due to limited attention of analysts and investors. We show that the underreaction to the text in these reports can systematically be exploited. An investment strategy that buys stocks or bonds of companies with few changes compared to previous reports ('non changers') and sells companies with many changes ('changers') over the period 1997-2017 earns significant alpha in the equity market, and high yield and investment grade corporate bond markets.

Keywords: corporate bonds, factor investing, SEC filings, stocks, text analysis

Table of contents

1 Introduction

Models for predicting asset prices generally only use numerical data, such as market prices of financial instruments and accounting data of companies. Such models thus ignore potentially relevant information that exists beyond these numerical data. An important source of non-numerical data is textual data since it accounts for most of all corporate information (Ur-Rahman & Harding, 2012). Examples of text data are news articles, social media posts, transcripts of management presentations, and corporate reports. Until about two decades ago, usage of such text sources required human intervention to code attributes into numerical form before any analysis could be conducted, which was a very slow and tedious process (Dörre, Gerstl, & Seiffert, 1999). Nowadays, due to advances in computer science and the immense growth in computing power, it is possible to use text mining techniques to systematically analyse vast amounts of text data (Kao & Poteet, 2007). This shift has lead to a staggering increase in studies that examine the benefits of enhancing quantitative financial models with text mining analytics.

One particular strand of research investigates the information content of the text in corporate reports that publicly listed companies in the U.S. have to file with the Securities and Exchange Commission (SEC) periodically or when a specific event happens, such as a merger or dept issuance. Amongst the corporate reports, most attention has historically been directed towards the annual (Form 10-K) and quarterly (Form 10-Q) reports. These reports unarguably are the primary source of company-specific information, because they provide an extensive overview of a business from an inside perspective.[1] The information that is present in 10-Ks and 10-Qs should allow anyone to fully grasp the state of a company. However, in practice, it is found that valuable information in the text in 10-Ks and 10-Qs is easily overlooked by analysts and investors simply because of the daunting challenge to read and grasp everything in the often more than 100 pages of text in these documents (Loughran & McDonald, 2014). This has paved the way for text mining approaches. So far, the results of text mining studies are encouraging as they have documented that certain characteristics of the text in 10-Ks and 10-Qs are predictive of stock price changes in the period after publication: the complexity of the text in terms of readability level and length of the text has been linked to stock price volatility (Bonsall, Leone, Miller, & Rennekamp, 2017), the sentiment of the text in terms of positive and negative words has also been linked to stock price volatility (Loughran & McDonald, 2011), and changes to the text in consecutive reports have even been linked to both stock price volatility and returns (Amel-Zadeh & Faasse, 2016; Cohen, Malloy, & Nguyen, 2018).

[1] The reports are so extensive because laws and regulations prohibit companies from making materially false or misleading statements and from omitting material information that is needed to make the disclosures not misleading.

Remarkably, almost all studies in this field have focused on predicting stock price movements even though corporate filings are also a good source of information for investors in other capital markets, such as the corporate bond market (Givoly, Hayn, & Katz, 2017). The reason for this lack of attention to the impact of the text in 10-Ks and 10-Qs on the corporate bond market is likely multifaceted. Part of it might be the lack of high quality corporate bond data until recently and the high level of difficulty in creating a good debt-to-equity link to tie bond prices and company fundamentals together. Another part might be that 10-Ks and 10-Qs are only filed by public companies while the corporate bond universe also contains a high fraction of private companies (Intertrust, 2018). Lastly, it might also play a role that quantitative investing in the corporate bond market has only started to gain traction in the last few years (Houweling & Van Zundert, 2017; Israel, Palhares, & Richardson, 2016), whereas it has already been established in the equity market for over two decades now (Carhart, 1997; Fama & French, 1993).

The purpose of this thesis is to bridge the gap by investigating the information content of the text in 10-Ks and 10-Qs with respect to both the stock and corporate bond market. To do so, we first download all 10-Ks and 10-Qs (in .txt format) published after 1994 from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR). Then, we link the filings to monthly price and accounting data of the 1,800 largest companies in the combined MSCI and BMI Developed Market indexes for stocks and to all public issuers in the Barclays U.S. Corporate Investment Grade (IG) and High Yield (HY) indexes for bonds. Next, we clean the downloaded 'complete submission files' with regular expressions such that all (numerical) tables, HTML elements, binary-to-text encodings (of .pdf, .xlsx, and .png files), and exhibits of other filings are removed and only the text, numbers, and symbols in the main body of the filing document remain. Finally, we (pre)process the text, construct readability, sentiment, and similarity variables, and perform a battery of tests.

The contribution of this thesis to the existing literature is fourfold. First, in a replication study on a more recent and larger sample of data than previously examined (which spans over two decades), we confirm the results from previous studies on the predictive power of readability, sentiment, and similarity variables for future stock price volatility and returns (Amel-Zadeh & Faasse, 2016; Bonsall et al., 2017; Cohen et al., 2018; Loughran & McDonald, 2011). In addition, we show that readability, sentiment, and similarity variables are highly uncorrelated and that they jointly predict stock price volatility even better than standalone. From this, we gather that the results for the stock market are robust.

Second, to the best of our knowledge, this is the first study that shows that readability, sentiment, and similarity variables constructed from the text in 10-Ks and 10-Qs also predict future corporate bond price volatility and returns. Moreover, we find that the signs of the effects are similar for stocks and bonds, which suggests that the stock and bond market are impacted in the same way by the text in 10-Ks and 10-Qs. Altogether, these results provide strong evidence that the text in 10-Ks and 10-Qs actually contains material information for capital markets and that previously documented findings for the stock market are not simply the result of data mining (Harvey, Liu, & Zhu, 2016).

Third, in the fashion of Fama and French (1993) we construct characteristics sorted portfolios based on readability, sentiment, and similarity variables and perform backtests to gauge whether the text variables can be used to build a profitable investment strategy. In line with Cohen et al. (2018), we find that the stocks of 'changers' significantly underperform those of 'non-changers'. In addition, we show that the strategy also works in the corporate bond market. The strategy earns significant annualized alpha's in excess of 3% for equities, 2% for HY, and 1% for IG, even when controlled for quantitative equity and bond factors. These results also hold across an array of robustness tests.

Fourth, we show that the performance of the similarity strategy in all asset classes is concentrated in the strongest downturn periods of the equity market. Strategies that outperform during crashes generate so-called crisis alpha, which makes them excellent diversifiers in a multi-strategy portfolio (Greyserman & Kaminski, 2014). To test whether this is also the case here, we add the composite similarity strategy to an equally-weighted multi-factor portfolio with size, value, and momentum and find that the Sharpe ratio significantly improves for all asset classes (t-stats > 3) due to a large decrease in volatility. Adding the similarity strategy thus robustly improves the risk-adjusted performance of the multi-factor strategy. This shows that including textual factors next to numerical factors in a multi-factor strategy can lead to more robust and better (risk-adjusted) performance.

The remainder of this thesis is organized as follows. Section 2 provides background information on the SEC, a detailed introduction of Forms 10-K and 10-Q, and an overview of the literature that has investigated the information content of the text in 10-Ks and 10-Qs. Section 3 introduces the panel regression framework, modern asset pricing theory, and the characteristics sorted portfolio methodology together with a host of performance measures. The data is presented in Section 4, and the text variables are constructed in Section 5. Section 6 presents and discusses the results. Section 7 concludes.

2 Background and related literature

2.1 Securities and Exchange Commission

During the Roaring Twenties (1920s), the U.S. economy and the stock market experienced a rapid expansion, and stocks hit record highs. In this period of post-war optimism, companies sold stocks and bonds on the promise of large profits but without disclosing any meaningful information to investors. This lead to a huge inflow of first-time investors that believed that the markets would continue to rise forever. Unfortunately for them, only a few years later, market sentiment shifted abruptly and the market lost more than half of its value in the Wall Street Crash of 1929. Following this market crash, the U.S. Congress enacted the Securities and Exchange Commission (SEC) to govern new federal securities laws that had to restore the public's faith in the capital markets: the Securities Act of 1933 and the Exchange Act of 1934.[2] These laws required companies to tell the public the truth about their businesses, the securities they are selling, and the risks involved in investing. Moreover, from then on, public companies were also required to annually report about their business operations, financial condition, and management in a comprehensive summary report labeled Form 10-K.

The founding of the SEC and the introduction of the 10-K were the first steps towards a better information flow to investors and a subsequently more efficient and transparent capital market system. Since then, the SEC has continually listened to the concerns of all major market participants and learned from their experiences, to further improve the information disclosure system. For example, in 1970, the SEC introduced a quarterly variant of the Form 10-K, the Form 10-Q, to provide investors with more timely information, and in 1994, the SEC adopted the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR), which performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies, to improve the efficiency and fairness of the distribution of time-sensitive corporate information.[3] Companies were phased in to EDGAR filing over a three-year period, ending May 6, 1996. As of that date, all publicly traded U.S. companies are required to submit their filings to EDGAR. Nowadays, there are over 150 types of forms that have to be filed with the SEC periodically or when a specific event happens, such as a merger or dept issuance.

2.2 Form 10-K and 10-Q

A company's 10-K is the primary source of company-specific information for investors. It is comparable but distinct from the annual report to shareholders that companies are required to publish before an annual meeting, since it covers similar subjects but in much more detail. Laws and regulations prohibit companies from making materially false or misleading statements in their 10-Ks. Likewise, companies are prohibited from omitting material information that is needed to make the disclosure not misleading. Therefore, the information that is present in a 10-K allows investors to fully grasp the state of a company.

The report is categorized into four major topics: i) The Business, ii) The Financials, iii) Executives, Directors and Corporate Relationships, and iv) Exhibits and Financial

[2] `https://www.sec.gov/Article/whatwedo.html`
[3] `https://www.sec.gov/edgar.shtml`

Statement Schedules. Within each topic, there are smaller subtopics (see Table 15 for the table of contents of a 10-K). Analysts mostly focus on Part ii, and more specifically, item 8, the financial statements section. This section contains tables with accounting numbers that describe the performance of a company, e.g. sales, earnings, and the amount of debt outstanding, for the current and past few years. These accounting numbers are used by investors to valuate companies based on financial ratios or 'multiples' such as the Book to Market, Debt to Equity, and Price to Earnings ratio (Koller, Goedhart, Wessels, et al., 2010). Moreover, shareholders want to maximize their wealth by investing in companies that have (long-term) growth potential. One way of gauging a company's ability to generate cash flow and earn profits in the future is by examining the progression of accounting numbers over time.

In addition to numeric information, there also is a large amount of unstructured textual information in 10-Ks. Take Bank of America's 10-K from 2017 as an example.[4] The 214-page report contains 6 pages of financial statements, about 25 pages of other tables, and the rest of the report consists of text. The text sections in 10-Ks are of a very technical nature by construction. Reading a complete 10-K filing thus would take a considerable amount of time. Therefore, investors are often advised to focus on a few key sections instead of reading the whole document.[5] Presumably the most important text section in the whole report is item 7, the Management's Discussion and Analysis section (MDA). In this section, the company's high-ranking officers analyze the company's performance and provide commentary on financial statements, compliance with laws and regulations, and actions it has planned or has taken to address any challenges the company is facing. Further, management also makes forward looking statements in which they discuss the upcoming year by outlining future opportunities and challenges. Another important text section, which was added in 2005, is item 1A, the Risk Factors section (RF). In this section, the SEC requires companies to discuss the most significant factors that make the company speculative or risky.

The SEC requires companies to file their 10-K within a prespecified window after the company's fiscal year end. The deadline is after 90 days for companies with less than \$75 million free float (non-accelerated filer), after 75 days for companies with \$75 to \$750 million free float (accelerated filer), and after 60 days for companies with more than \$750 million free float (large accelerated filer). If a company fails to file their 10-K in time, then they have to file a Non-Timely (NT) 10-K notification, which grants them the right to a 15 day extension of the submission deadline. If the extended deadline is missed, consequences may include loss of SEC registration, de-listing from stock exchanges as well

[4] `https://www.sec.gov/Archives/edgar/data/70858/000007085817000013/bac-1231201610xk.htm`

[5] See e.g. `http://www.rationalwalk.com/?p=15643` and `https://www.investopedia.com/articles/basics/10/efficiently-read-annual-report.asp`

as possible legal consequences. Late filings are usually caused by internal problems related to the financial well-being of a company. Therefore, filing an NT 10-K is perceived as very negative by the market (Bartov & Konchitchki, 2017).

Next to an annual report, companies also have to file a Form 10-Q, which is a quarterly report. The 10-Q is an unaudited version of the 10-K that covers less topics and is less detailed (see Table 15 for the table of contents of a 10-Q). A 10-Q has to be filed three times a year, after the first, second, and third fiscal quarter end, with a deadline after 45 days for non-accelerated filers, and after 40 days for (large) accelerated filers. Analogously, an NT 10-Q notification has to be filed when a company fails to file their 10-Q in time.

2.3 Literature review

In the last decades, computing power has increased exponentially and the fields of Machine Learning (ML) and Natural Language Processing (NLP) have made spectacular progress. This has lead to a growing literature that applies text analysis to financial applications, such as analyzing corporate filings.

2.3.1 Text analysis of corporate filings

Prior research on the analysis of text in corporate filings can be divided into three main strands: i) readability, ii) sentiment analysis, and iii) similarity.

Readability of a text is important since the difficulty of a text directly affects the degree to which people are able to comprehend it. A popular readability measure from computational linguistics is the Fog Index, which is a function of average sentence length and percentage of complex words (Gunning, 1952). The measure predicts the minimum grade level needed to understand a text. Lower levels of this index thus indicate more readable text. Li (2008) uses the Fog Index to measure the readability of the MDA in 10-Ks. He finds no relation between readability and future stock returns. As a follow up, Loughran and McDonald (2014) show that the Fog index is poorly specified in a financial context.[6] They argue that the focus should be on information complexity rather than readability. They use the total file size of a filing as a proxy for the complexity and find that firms with larger 10-K file sizes have significantly higher stock price volatility. Hence, they conclude that shorter filings are more likely to be read, which leads to better incorporation of the information into stock prices. Bonsall et al. (2017) show that the variation in filing size over time is greatly driven by changes in unrelated content to the actual text in the

[6]The Fog index defines complex words as words containing more than two syllables. However, words such as *financial*, *company*, *operations*, *management*, *employees*, and *customers* are typically not considered to be complex for investors. Moreover, Loughran and McDonald also document that the sentence extraction method of Li fails in more than 50% of the cases because measuring sentence length in the context of financial disclosures is substantially less precise than measuring sentence length in traditional prose.

10-K (e.g., HTML, XML, .pdf and .jpeg file attachtments). Therefore, they suggest to use the length of the actual text in the filing as a proxy for the complexity. Amel-Zadeh and Faasse (2016) confirm that longer documents indeed lead to higher volatility in the period after publication.

Another important characteristic of a text is its tone or sentiment. In a seminal paper, Tetlock (2007) finds that the proportion of negative words in the 'Abreast of the Market' column in the Wall Street Journal robustly predicts downward pressure on stock market prices. He shows that the stock market incorporates the information embedded in negative words with a slight delay, which means that a short-term trading strategy can be constructed. In the paper, he classifies words as pessimistic or negative if they appear in the negative wordlist of the Harvard Psychosociological Dictionary. Li (2010) applies the same approach to the MDA section of 10-Ks and finds that net tone (difference between positive and negative words) is not (positively) associated with future performance. He concludes that the dictionary based approach might not work well for corporate filings. In response, Loughran and McDonald (2011) show that almost three-fourths of the negative words in the Harvard dictionary are typically not negative in a financial context.[7] Therefore, they create a new dictionary that is focused on finance and classifies words in six classes (positive, negative, uncertainty, litigious, strong modal, and weak modal). They apply their dictionary to the MDA of 10-Ks and find that the sentiment variables predict stock price movements and company events in a short window around the filing date. Kearney and Liu (2014) report that the Loughran and McDonald wordlists have become predominant in studies on sentiment analysis in finance. Although widely used, the wordlists are not exhaustive. For instance, Karapandza (2016) constructs a future tense wordlist, which consists of words that indicate forward looking statements, and finds that firms that talk less about the future in their 10-K filing have higher volatility than firms that talk more about the future.

Finally, firms usually tend to repeat what they most recently reported (Cohen et al., 2018). Therefore, a change in a text is informative because of the new content itself, and due to the fact that a decision was made to make a change. Changes in content can for example be captured by filing-over-filing changes of readability and sentiment variables. Chouliaras (2015) finds that yearly changes of sentiment variables are better able to predict market movements than the levels themselves. Correspondingly, Amel-Zadeh and Faasse (2016) find that investors generally underreact to incremental negative information in the MDA section of 10-Ks as it is predictive of negative future stock returns (up to one year), whereas the level of negative information has no predictive power. To capture

[7] Words that are typically not negative in a financial setting, which are included in the Harvard negative wordlist, are, e.g. *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*.

the magnitude of a change, Brown and Tucker (2011) compute a document modification score that looks at year-over-year changes of a firms' MDA section in the 10-K filing. This modification score is based on the cosine similarity measure, which is a commonly used measure from the information retrieval literature (Singhal et al., 2001). They show that this modification score has declined considerably, since MDAs have quickly gotten larger over the years. Further, they find no relation between modification scores and stock prices. Cohen et al. (2018) compute four similarity scores (based on cosine, Jaccard, and simple similarity, and minimum edit distance) and apply them to the text in (subsections of) 10-Ks and 10-Qs. In contrast to Brown and Tucker, they find that 'changers' consistently underperform 'non-changers'. Moreover, they report that they are able to construct a hedge portfolio based on similarity scores that earns significant (risk-adjusted) returns even when controlled for other known risk factors.

2.3.2 Corporate bonds

Most previous research on the analysis of 10-Ks and 10-Qs has focused on predicting stock price movements, even though corporate filings are also a good source of information for other capital market participants, such as creditors, who invest in the corporate bond of a company (Givoly et al., 2017). A corporate bond is a certificate of debt issued by a company that guarantees payment of the borrowed amount (the par value) plus interest (the yield) by a specified future date (the maturity date), if the company has not defaulted by then (Fabozzi, Pollack, & Fabozzi, 1995). All bonds carry a rating that indicates the creditworthiness of the bond's issuer. Credit ratings are assigned by three independent rating agencies: Moody's, Standard & Poor's, and Fitch. Bonds with a rating of BBB- (on the Standard & Poor's and Fitch scale) or Baa3 (on Moody's) or better are considered 'investment grade' (IG), while bonds with lower ratings are considered speculative and often referred to as 'junk' or 'high yield' (HY) bonds. Lower-rated bonds generally offer higher yields to compensate investors for the additional (default) risk. The two corporate bond segments have vastly different risk and return profiles and therefore they are generally seen as different asset classes by practitioners.

Last year, the first papers were published that consider the information content of text in corporate filings for credits. Ertugrul, Lei, Qiu, and Wan (2017) examine the effect of readability and tone (based on the uncertainty and weak modal wordlists of Loughran and McDonald) of text in 10-Ks on bank loan contract terms. They find that firms with longer (less readable) annual reports, face higher loan spreads and more restrictive contract terms, such as shorter maturity and a greater likelihood of collateral requirements. Moreover, they find that an ambiguous tone in annual reports results in higher loan spreads and greater price crash risk, which suggests that managers might use ambiguous language to cover up adverse news. Bonsall et al. (2017) find that less readable 10-Ks are associated with less favorable ratings, greater bond rating agency disagreement, and a higher cost of debt. Chin, Liu, and Moffitt (2018) introduce a novel credit specific methodology. They examine whether the ordering of risk factors in the RF in 10-Ks is related to the debt terms of a firm. They find that firms that place the credit risk disclosure closer to the beginning have lower credit ratings, higher bond spreads, and are also more likely to experience rating downgrades and bankruptcies in the future.

3 Methodology

This section describes the methods that we use in this thesis to investigate the informativeness of the text in 10-Ks and 10-Qs for stock and corporate bond markets. First, Section 3.1 introduces the panel regression framework, whereafter Section 3.2 visits modern asset pricing theory, characteristics sorted portfolios, and performance evaluation measures.

3.1 Regression analysis

The data (both text and financial data) that we use in this thesis is collected over two dimensions, namely individuals (companies) and time (months). To benefit from the information in both dimensions we perform a generalized version of standard regression known as panel regression (Heij et al., 2004). A linear panel data model is given by

$$y_{it} = c + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T, \tag{1}$$

where N denotes the size of the cross-section, T denotes the length of the time-series, y_{it} is the dependent variable, c is a constant, $\boldsymbol{x}_{it}$ is a $(K \times 1)$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of parameter coefficients, and u_{it} is an error term. The parameters in this model can unbiasedly and consistently be estimated by Ordinary Least Squares (OLS) when all observations are independent and identically distributed. The content of the text in all 10-Ks and 10-Qs in the panel is not likely to be independent, however. First, companies in the same industry are closely related and face similar risks and opportunities, which means that their reports will be very much alike. Second, the SEC changes the requirements for the reports (almost) every year, which affects the content of the reports of all companies.

3.1.1 Two-way fixed effects model

The two-way fixed effects regression model (Wooldridge, 2010) accounts for unobservable individual- and time-specific fixed effects by extending the specification in (1) with an individual-specific intercept μ_i and time-specific intercept λ_t as follows

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} \qquad\qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{2}$$

$$u_{it} = \mu_i + \lambda_t + \nu_{it}, \tag{3}$$

where ν_{it} is an independent and identically distributed error term with mean zero and variance σ_ν^2.

We assume that the individual-specific intercepts for observations of companies in the same industry are similar, such that $\mu_i = \mu_g$ if company i is in industry g for $i = 1, \ldots, N$ and $g \in \{1, \ldots, G\}$, where $\mu^{(g)}$ is the intercept for industry g and G is the total number of industries in the cross-section, even so as that the time-specific intercepts for observations from the same year are similar, such that $\lambda_t = \lambda^{(h)}$ if time period t is in year h for $t = 1, \ldots, T$ and $h \in \{1, \ldots, H\}$, where μ_h is the intercept for year h and H is the total number of years spanned by the time-series. This reduces the model to the following form

$$y_{itgh} = \boldsymbol{x}_{itgh}'\boldsymbol{\beta} + \mu^{(g)} + \lambda^{(h)} + \nu_{itgh} \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \quad g \in \{1, \ldots, G\}, \quad h \in \{1, \ldots, H\}, \tag{4}$$

where the subscripts g and h denote the industry and year groups of observation (i, t).

3.1.2 Two-way fixed effects estimator

The model can be written in matrix form as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\lambda} + \boldsymbol{\nu}, \tag{5}$$

where $\mathbf{y}$ is a $(NT \times 1)$ vector, $\mathbf{X}$ is a $(NT \times K)$ matrix, and $\boldsymbol{\nu}$ is a $(NT \times 1)$ vector which are obtained by first grouping observations by company and industry and then stacking them over all industries, $\boldsymbol{\mu} = [\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(G)}]'$, $\boldsymbol{\lambda} = [\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(H)}]'$, and $\mathbf{D}$ and $\mathbf{Z}$ are $(NT \times G)$ and $(NT \times H)$ indicator matrices with $\mathbf{D}_{jg}$ equal to 1 if observation j is from a company in industry g and zero otherwise and $\mathbf{Z}_{jh}$ equal to 1 if observation j is from year h and zero otherwise. This model is commonly referred to as the least squares dummy variable (LSDV) model. Although theoretically sound, this implementation is not very sensible in practice, since it involves calculating the inverse of a $(G+H+K) \times (G+H+K)$ matrix which is numerically cumbersome when $G+H$ is large. Fortunately, the model can be rewritten into an equivalent form which is more efficient in terms of computation.

The Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963) states that the OLS estimate of the LSDV model is equivalent to the result from a regression of $\widetilde{\mathbf{y}} = \mathbf{M_S}\mathbf{y}$ on $\widetilde{\mathbf{X}} = \mathbf{M_S}\mathbf{X}$, where $\mathbf{M_S} = \mathbf{I} - \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$ is the annihilator matrix of $\boldsymbol{S} = [\boldsymbol{D}, \boldsymbol{Z}]$.[8]

[8]Premultiplying with the annihilator matrix effectively removes the fixed effects from all observations

Hence, the OLS estimator is given by

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}}). \tag{6}$$

3.1.3 Two-way cluster-robust variance estimator

Common shocks to companies in the same industry or to all companies in a given year could make error terms correlated. Standard robust variance estimators do not take this correlation into account, which means that standard errors would likely be underestimated (Abadie, Athey, Imbens, & Wooldridge, 2017). Cameron, Gelbach, and Miller (2011) introduce multi-way cluster-robust standard errors that control for possible correlations within clusters. Assuming that the errors are uncorrelated between observations from different industries or from different years such that

$$\mathrm{E}[\nu_{itgh}\nu_{itg'h'}|\widetilde{\mathbf{x}}_{itgh}, \widetilde{\mathbf{x}}_{itg'h'}] = 0 \text{ unless } g = g' \text{ or } h = h', \tag{7}$$

and that errors for observations from the same industry or from the same year may be correlated with quite general heteroskedasticity and correlation. Then, the variance of the parameter estimates can consistently be estimated by the two-way cluster-robust variance matrix

$$\widehat{\mathrm{V}}[\widehat{\beta}] = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widehat{\mathbf{B}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}, \tag{8}$$

as $\min\{G, H\} \to \infty$, with

$$\widehat{\mathbf{B}} = \widetilde{\mathbf{X}}(\widehat{\boldsymbol{\nu}}\widehat{\boldsymbol{\nu}}' \circ \boldsymbol{S}^{NT})\widetilde{\mathbf{X}}, \tag{9}$$

where $\widehat{\boldsymbol{\nu}} = \widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}$, $\boldsymbol{S}^{NT}$ is a $NT \times NT$ indicator matrix with ijth entry equal to 1 if the ith and jth observation are from the same industry and/or from the same year and equal to zero otherwise, and $\circ$ denotes element-by-element multiplication. In practice, finite-sample modifications of this variance estimator are typically employed, since without modification the cluster-robust standard errors are biased downward in finite samples. We use the modification proposed by Cameron et al. (2011) to offset the deflation.[9]

3.2 Investment analysis

The efficient market hypothesis (EMH) postulates that newly revealed information about a company is (almost) immediately reflected in its stock price (Fama, 1965). In weak-form

in Z by demeaning, since $\widetilde{z}_{itjs} = z_{itjs} - \bar{z}_j - \bar{z}_s + \bar{z}_{js}$ with $\bar{z}_k$ the mean over all observations in group k and $\bar{z}_{kl}$ the mean over all observations in groups k and l for a $(NT \times 1)$ vector Z with individual group dimension j and time group dimension s.

[9] To improve the finite sample properties of the cluster-robust standard errors the estimated residuals are multiplied by $\sqrt{c}$, where

$$c = \frac{H}{H-1}\frac{NT-1}{NT-K}.$$

efficiency, this covers all historical price data. In semi-weak-form efficiency, this covers all historical price data and publicly available information. Finally, in strong-form efficiency, this covers all information, be it public or private. A direct implication of the EMH is that future returns are random, which means that it is impossible to consistently beat the market. A large body of empirical evidence, however, claims that returns are predictable (up to a certain degree) and that it should be possible to consistently beat the market.

3.2.1 Capital Asset Pricing Model

The capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) states that the expected return $\mathrm{E}[R_i]$ of a company's stock is wholly determined by its exposure to the expected market risk premium or excess market return (RMRF), such that

$$\mathrm{E}[R_i] = \beta_i \mathrm{E}[RMRF], \tag{10}$$

where $\beta_i = \sigma_{i,m}/\sigma_i^2$ is the market beta of company i. Here, $\sigma_{i,m}$ is the covariance between the stock returns of company i and market returns and σ_i^2 is the variance of the stock returns of company i. The CAPM thus implies that risky stocks will have higher returns than less risky stocks. A way to actually test the strength of the relation between a stock's return and the market premium is the following time-series regression

$$R_{i,t} = \alpha + \beta_i RMRF_t + \varepsilon_t, \tag{11}$$

where α is generally referred to as the CAPM alpha and ε_t is a standard normal error term. If $\alpha = 0$ then the stock's return is fully explained by the market premium, whereas if $\alpha \neq 0$ then there supposedly are additional factors driving the stock's return. Empirically, it is found that the linear relation in (10) does not hold (so $\alpha \neq 0$). In fact, there is evidence that the relation between market beta and return might even be inverted (Blitz & Van Vliet, 2007).

3.2.2 Fama French and Carhart factor models

Fama and French (1993) extend the CAPM with two factors: Small Minus Big (SMB) which is the historical return of small caps over large caps, and High Minus Low (HML) which is the historical return of value companies (high book-to-market ratio) over growth companies (low book-to-market ratio). The Fama-French three-factor model is defined as

$$R_{i,t} = \alpha^{3F} + \beta_1 RMRF_t + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_t, \tag{12}$$

where α^{3F} is the three-factor alpha, and β_2 and β_3 respectively are the sensitivities of returns to the HML and SMB factors. Moreover, Jegadeesh and Titman (1993) find that stocks that perform the best (worst) over a three to 12 month period (excluding last month's return because of the well documented one-month reversal in stock returns) tend to continue to perform well (poor) over the subsequent months. Carhart (1997) further explores this effect and finds similar results. He extends the Fama-French three-factor model with the Up Minus Down (UMD) factor which is the historical return of winners over losers over the previous 12 months excluding last month's return. The Carhart four-factor model is defined as

$$R_{i,t} = \alpha^{4F} + \beta_1 RMRF_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 UMD_t + \varepsilon_t, \tag{13}$$

where α^{4F} is the four-factor alpha and β_4 is the sensitivity of returns to the UMD factor.

The SMB, HML, and UMD factors have performed very well over long samples and in various different countries and industries (Dimson, Marsh, & Staunton, 2017). There are different theories that try to explain the stellar historical performance of these factors (Ang, 2014). One view is that factors earn a premium as a compensation for bearing systematic risk. A second view is that factors earn a premium by benefiting from systematic errors in the market that arise either because investors exhibit behavioral biases (e.g. overconfidence, loss aversion) or because they are subject to certain constraints (e.g. investment horizon, no short selling). The literature is not conclusive on the (in)correctness of either view.

In the meantime, hundreds of new factors have been discovered that try to explain the cross-section of expected stock returns—the so-called factor zoo (Cochrane, 2011). In-sample, all these factors are highly significant. Out-of-sample (e.g. in a different time period or country), however, only very few actually perform well (Harvey et al., 2016). Therefore, most of the factors in the factor zoo do not seem to capture truly new dimensions of risk. Instead, they are more likely to be statistical flukes. This emerging problem of data mining underscores the importance of good robustness testing. One particularly strong robustness test is to perform the same analysis on another asset class.

3.2.3 Factor models in the corporate bond market

The literature on factor models in the corporate bond market has only developed quite recently. Almost all papers on the subject have been published in the (latter part of the) last decade. Only a few years ago, Houweling and Van Zundert (2017) showed that all factors from the Carhart four-factor model are also present in the corporate bond market. They define the variables as follows:

- The market beta is proxied by the DTS beta, which is defined as

$$\beta^{DTS}_{i,t} = \frac{DTS_{i,t}}{DTS^m_t} \tag{14}$$

for company i at time t. Here $DTS_t^m = N^{-1}\sum_{j=1}^N DTS_{j,t}$ where N is the size of the universe.

- Size is defined as the total index weight of each company, calculated as the sum of the market value weights of all its bonds in the index.
- Value is defined as the percentage difference between the actual credit spread and a fitted credit spread. The fitted ('fair') credit spread is estimated by the following cross-sectional regression

$$S_i = c + \sum_{r=1}^{21} \beta_r \mathbf{1}_{\{ir\}} + \gamma M_i + \delta \Delta S_i + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, N, \tag{15}$$

  where S_i is the spread of bond i, $\mathbf{1}_{\{ir\}}$ is equal to 1 if bond i has rating r, and 0 otherwise, M_i is the maturity of bond i, and ΔS_i is the 3-month change in the spread.
- Momentum is defined as the corporate bond return over the previous six months with a one month implementation lag.

An important design choice here is that all factors only use bond characteristics, such as rating, maturity, and credit spread, and no accounting data or equity market information. This ensures that all bonds can be included, and not only bonds with publicly listed equity. Moreover, the factors are created separately for the HY and IG market to follow the convention of treating these segments of the corporate bond market as two different asset classes.

3.2.4 Characteristics sorted portfolios

A popular method in empirical finance to test for predictive power of future stock or bond returns is to perform historical backtests with characteristics-sorted portfolios (Patton & Timmermann, 2008). For this method, every month all companies in the cross-section are sorted in descending (ascending) order on a given characteristic (e.g. the market cap of a company or the number of negative words in a 10-K or 10-Q) such that the first ranked company has the highest (lowest) exposure to the characteristic, the second ranked company has the second to highest (lowest) exposure to the characteristic, and so on and so forth. Then, in each month, the companies are split into J roughly equal disjoint groups based on their rank, such that the companies with the $\lfloor N/J \rfloor$ highest ranks are in the top group, the companies with a rank between $\lceil N/J \rceil$ and $\lfloor 2N/J \rfloor$ are in the second group, et cetera, where $\lfloor \cdot \rfloor$ is the floor operator and $\lceil \cdot \rceil$ the ceiling operator. In the literature, J is usually set to 5 (or 10 if the resulting groups are large enough, i.e. larger than 25). The groups are then turned into portfolios where each company is either given an equal weight (equally-weighted) or a weight proportional to its market weight (market-weighted). Next, in each time period, the stocks or bonds of the companies in each portfolio are held for H months. If $H = 1$, then only one portfolio is held at each point in time. When $H > 1$, then multiple portfolios are held at the same time according to the overlapping-portfolio methodology introduced by Jegadeesh and Titman (1993). Finally, long-only portfolios are constructed by separately investing 1 long dollar in each of the J portfolios, and a hedge or top-minus-bottom (TB) portfolio is constructed by investing 1 long dollar in the top portfolio (the long leg) and 1 short dollar in the bottom portfolio (the short leg).

3.2.5 Performance evaluation

There are many different measures that can be used to evaluate the performance of an investment strategy. The importance of these measures differs widely amongst investors because of differences in investment objectives. Therefore, we consider multiple types of measures to evaluate the performance from as broad a perspective as possible.

Return statistics are absolute measures that allow an investor to assess the profitability and riskiness of a strategy. Examples are portfolio return, volatility, and Sharpe ratio (risk-adjusted return). On their own, these statistics portray much information about the performance of a portfolio, however, often an investor is also concerned with the portfolio's performance relative to another portfolio, e.g. the market portfolio.

Outperformance statistics are relative measures that allow an investor to assess the performance and riskiness of a strategy relative to a benchmark. Examples are outperformance (difference in return), tracking error (volatility of outperformance), and information ratio (outperformance over tracking error). Moreover, to be able to judge the difference in performance in a statistical sense, we perform hypothesis tests on the difference in return, variance, and Sharpe ratio of two portfolios. The most popular test, amongst practitioners and empirical researchers, for comparing Sharpe ratios is the test of Jobson and Korkie (1981) and its corrected version by Memmel (2003). However, this test is not valid when returns have tails heavier than the normal distribution or when returns are of a time-series nature. Both features apply here, and therefore we use the test of Ledoit and Wolf (2008) instead, which is based on robust inference methods. This test also applies to differences in variance when a slightly adjusted version is used (Ledoit & Wolf, 2011). For a detailed explanation of these hypothesis tests, see Appendix B.

Lastly, we consider alpha statistics, to assess the 'economic significance' of the performance of a portfolio. That means performance in excess of, for example, the market portfolio (CAPM alpha) or all the factors in the Carhart four-factor model (four-factor alpha) in a time series regression setting.

4 Data

4.1 Equity & corporate bond data

We obtain equity data from FactSet for all stocks of companies that are amongst the 1,800 largest U.S. companies in a month, based on market value, in the combined MSCI and BMI Developed Market indexes. In this way, most nano- and micro caps are filtered out, which ensures that (il)liquidity will not be a (major) driver of any of the findings (Ibbotson & Kim, 2017). The sample starts in 1994M1, since from then on corporate filings are available from the EDGAR database, and ends in 2017M12 (276 monthly observations). For each stock in each month, various characteristics are provided by FactSet such as market value, book-to-market ratio, share price, and return. To focus on the equity premium, we convert the returns to excess returns by subtracting the yield on 10-year U.S. Treasuries.

We collect corporate bond data, over the same sample period, from Barclays for all bonds of issuers in the Barclays U.S. Corporate Investment Grade and High Yield Indexes. For each bond in each month various characteristics are provided by Barclays such as age, amount outstanding, seniority, time to maturity, credit rating, credit spread, duration, bond price, and return. Based on duration and spread we construct Duration Times Spread (DTS) which is the market standard method for measuring the credit volatility of a corporate bond (Dor et al., 2007). Further, we convert the credit rating variable to a numerical scale (AAA = 1, AA+ = 2, ...) such that it can be aggregated and used in my analysis. The data is provided at the bond level, however, we are interested in the issuers of bonds and not the bonds themselves. Therefore, we aggregate all bond level data to issuer level data by only selecting the "most representative bond" of an issuer in each month (Haesen, Houweling, & Van Zundert, 2013), where the most representative bond is selected with a filter based on (i) seniority, (ii) maturity, (iii) age and (iv) amount outstanding.[10] The resulting sample of bonds is relatively liquid. Hence, there should be no concerns about exploiting a possible (il)liquidity premium associated with investing in illiquid bonds (Palhares & Richardson, 2018). To focus on the credit premium (default premium), we convert the returns to excess returns by subtracting the yield on duration-matched 10-year U.S. Treasuries (duration matching removes the term premium that is present in the returns).

4.2 Text data

For our analysis, we want to obtain all 10-Ks and 10-Qs filed by the companies in the equity and corporate bond samples from 1994 to 2017. These filings are available from EDGAR and can be accessed through EDGAR's file system by looking up the filed documents

[10] Results are robust to different aggregation schemes (untabulated), e.g. market-value weighted.

associated with the Central Index Key (CIK) of the respective companies. Unfortunately, there is no readily available matching between the companies in my samples and EDGAR's CIK identifiers. Therefore, we have to link the databases myself. To do this, we develop a company name matching algorithm based on fuzzy string matching.[11] After matching the companies, we download all 10-Ks, 10-K405s[12] and 10-Qs filed by the matched companies from EDGAR, using a file scraper written in the Julia programming language. The filings are downloaded as 'complete submission files'. A complete submission file is a plain text file that contains the actual text in the filing document, but also a header with company and filing metadata, (numerical) tables, HTML elements, binary-to-text encodings (of .pdf, .xlsx, and .png files), and exhibits of other filings. First, we extract the header (the part between <SEC-HEADER> and </SEC-HEADER> tags) from the text file and parse it to obtain the filing's metadata: the company's name, CIK, Standard Industrial Classification (SIC), state of incorporation, the filing date and time, and the company's fiscal year or quarter end date. Then, we also extract the filing document (the part between <DOCUMENT> and </DOCUMENT> tags) from the text file and 'clean' it with regular expressions, such that only the text, numbers, and symbols in the main body of the original filing remain. Finally, we extract the MDA and RF sections from the filing document with a text parser which is also based on regular expressions.[13] We require at least 250 words to appear in the MDA and RF sections, because in many cases this information is 'incorporated by reference'. Meaning that the information can be found in an earlier published company report, typically the shareholders annual report. Moreover, 'smaller reporting companies', which are companies with a public float of less than $250 million or annual revenue of less than $100 million, are sometimes allowed to skip certain sections.[14]

4.3 Descriptive statistics

Descriptive statistics for the equally-weighted equity, HY, and IG markets of the complete sample ('All') and the matched sample ('Matched') are shown in Table 1. On average, the equity, HY, and IG samples contain 1,800, 481, and 513 companies, while the matched samples contain 1,693, 394, and 440 companies, which means an average coverage of 94%, 82%, and 86%, respectively. Given the difficulty in matching companies over a period

[11] The name matching algorithm uses character-level n-grams and TF-IDF to match the databases with high accuracy, even when the name of a company is stored differently in the two databases. Checks are performed to filter out false matches. A substantial effort is done to match these companies manually.

[12] Prior to 2003, a 10-K405 had to be filed (instead of a 10-K) when a company's personnel had not disclosed their insider trading activities within the required time period. Other than a marked checkbox on the cover page, there is no difference in the substance of the 10-K and 10-K405.

[13] The parser splits a filing into its subsections even when some subsections are missing, when the subsections are referenced from other places in the report, or when the headings contain typos or artifacts. This is successful in most cases. However, making it robust for filings from the earlier parts of the sample is especially hard because of the many inconsistencies in the reports.

[14] `https://www.sec.gov/smallbusiness/goingpublic/SRC`

of more than two decades, achieving a coverage greater than 80% for all asset classes is quite satisfying. To further evaluate the quality of the match between the databases we consider the coverage of the markets over time. For equities, the match is slightly better at the end than at the start of the sample period, while for HY and IG, the coverage is very stable throughout. Even though there is no perfect coverage, characteristics, such as return and volatility for all samples, and maturity and Duration Times Spread (DTS) for the corporate bond samples, of the original and matched samples are almost identical. This implies that the unmatched companies do not differ from the matched companies in a systematic way. Hence, the matched sample is representative of the original sample, which means that conclusions derived for the matched sample should hold more generally.

Table 1: Descriptive statistics of equity and corporate bond data

	Equity		HY		IG	
	All	Matched	All	Matched	All	Matched
Return statistics						
Mean return	9.16%	9.08%	3.33%	3.27%	0.83%	0.83%
Volatility	18.10%	18.15%	10.14%	10.31%	4.38%	4.45%
Sharpe ratio	0.51	0.50	0.33	0.32	0.19	0.19
Credit-specific statistics						
Maturity			7.08	6.99	10.56	10.57
Spread			533	532	171	169
Rating			14.55	14.56	7.71	7.75
DTS			2,368	2,364	1,124	1,117
Total sample	5,945	5,594	2,215	1,863	1,390	1,249
Average sample	1,800	1,693	481	394	513	440
Coverage		0.94		0.82		0.86

This table shows descriptive statistics for the equity, HY, and IG market of the initial complete sample ('All') and the matched sample ('Matched'). The statistics are first computed in the cross-section and then averaged over the time-series (1994M1-2017M12).

From here on we focus on the matched samples only. The return statistics for the three asset classes clearly show a difference in risk and return characteristics. The equity market is most risky (18.15% volatility) and has the highest returns (9.08%), the IG market is safest (4.45% volatility) and has the lowest returns (0.83%), and the HY market lies in between (10.32% volatility and 3.27% return). In terms of risk-adjusted returns (Sharpe ratio) the equity market (0.51) has been the most attractive asset class over the sample period, the IG market least attractive (0.19), and the HY market in between the others (0.32). Indicating a positive risk-return trade-off between asset classes, meaning that asset classes with higher risk are rewarded with even more return. The difference in risk and return between the HY and IG markets can be attributed to the different types of bonds that the markets are composed of. The HY market consists of bonds with a low maturity (7.1 years), high credit spread (532 bps), low rating (14.6), and high DTS (2,368), whereas the IG market consists of bonds with a high maturity (10.6 years), low credit spread (169 bps), high rating (7.8), and low DTS (1,117).

Figure 1 shows the log cumulative returns for the three markets. All three markets have gone up over the sample period with some downturns in between. The three most severe equity market downturns, in the last 20 years, are highlighted with grey bars, namely the Dot-Com bubble of 2000-2002, the Great Financial Crisis of 2008, and the stock market crash of 2016. In downturns, the equity and credit markets seem to move down together, and in upswings, they seem to go up together. This is reflected in the correlations between the markets, as the correlations between returns of the equity market and HY and IG markets are 64% and 56%, respectively, and the correlation between returns of HY and IG markets is 88%.

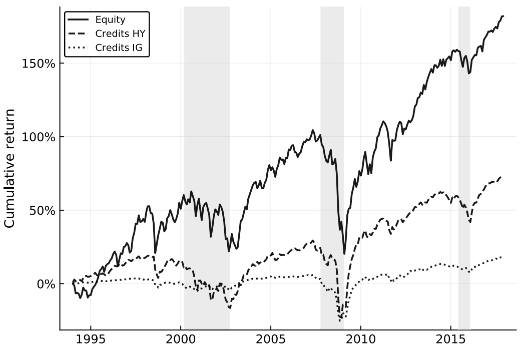


Fig 1: Equally-weighted market returns

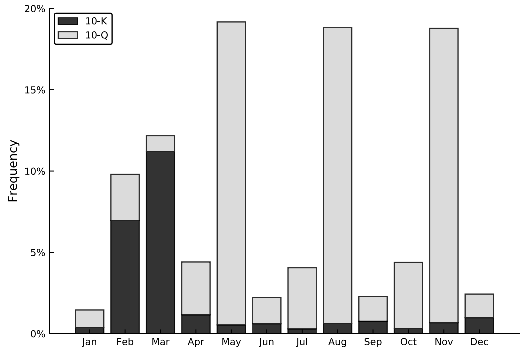


Fig 2: Monthly file date distribution

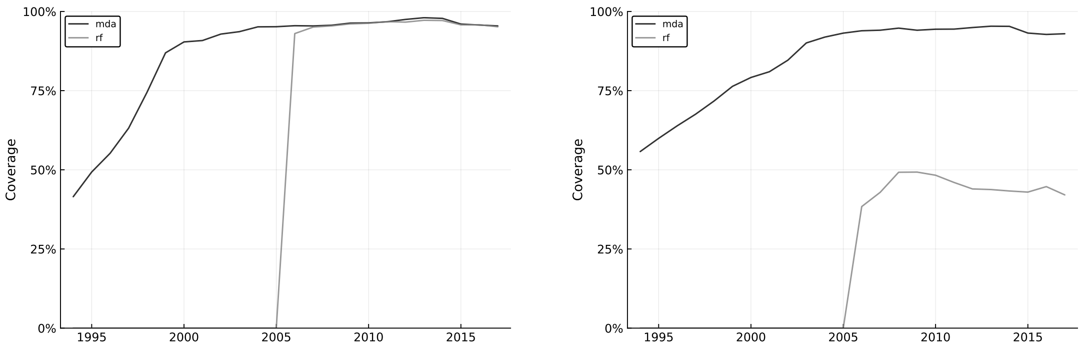
Table 2 shows the combined size of the filing data for the equity, HY, and IG samples. The number of 10-Ks in the sample is 55,302, and the number of 10-Qs in the sample is 169,045. Together, this results in a total sample of 224,347 filings. Figure 2 shows the distribution of file dates for 10-Ks and 10-Qs over the year. Most companies have December 31 as their fiscal year end, and consequently, most 10-Ks are filed in February and March, and most 10-Qs in May, August, and November. In the remaining months, the number of new filings is relatively low. Therefore, the filing dataset is unbalanced by construction, with many observations in the former months and only few observations in the latter months. Moreover, a company only files a new 10-K or 10-Q every quarter, which means that there are at most four data points per company per year.

The total number of MDAs in the sample is 195,348, which means that for 14% of the filings we have not been able to obtain a MDA. This decrease is mainly due to the exclusion of MDAs that are incorporated by reference, but also due to some MDAs that could not successfully be extracted. Most often, extraction failed either because a MDA was not included or because a firm uses an irregular or inconsistent filing template. The total number of RFs in the sample is 65,575, which is significantly less than the sample of complete filings after January 2007. The main reason for the difference is that RFs are relatively often included by reference in 10-Qs.

Table 2: Total number of filings and succesfully extracted MDAs and RFs

	10-K	10-Q	Total
Filing	55,302	169,045	224,347
MDA	49,544	145,804	195,348
RF	27,612	37,963	65,575

Figure 3 shows the coverage over time for MDAs and RFs in 10-K and 10-Q filings. At the start of the sample, almost half of the filings have MDAs that are incorporated by reference or that can not successfully be extracted. Over time, this rate quickly decreases, and towards the end, the exclusion rate drops below 10%. The initial high levels of exclusion can be attributed to a somewhat soft regulation policy by the SEC and the firms being inexperienced with the filing process. Over time, both have improved, and nowadays less MDAs are included by reference and filings have become more consistent and easier to process. The exclusion rate of RFs is constant over time and comparable to the exclusion rate of the MDA for 10-Ks but much lower for 10-Qs. Once again, this is caused by the high number of RFs included by reference in 10-Qs.



(a) 10-K **(b)** 10-Q

Fig 3: Coverage of MDA and RF sections over time

5 Text analysis

In this section, we perform text analysis on 10-Ks and 10-Qs to construct variables that will be used as inputs for the quantitative analysis in the remainder of this thesis.

5.1 Text preprocessing

Humans are able to 'see structure' and spot word boundaries, punctuation marks, and other syntactical features of written text. In contrast, computers see text as just a collection of characters, numbers, symbols, and white spaces (Miner et al., 2012).[15] Therefore, to make

[15] Altough, the gap between humans and computers is diminishing rapidly. In certain limited areas deep learning models have achieved and even surpassed human level performance. An example is BERT,

text understandable for a computer, it first needs to be converted to a more structured data format. A major part of this conversion to structured data is cleaning or preprocessing the unstructured data. The preprocessing procedure that we follow to clean the text consist of tokenization, accounting for negations, and removing stopwords.

5.1.1 Tokenization

The first preprocessing task is to split the text in a document into individual words or 'tokens' and stack them in an ordered list. While this might seem like a simple exercise to a human, it has proven to be very difficult for computers, since they have no good understanding of what constitutes a word, sentence, or paragraph (Grefenstette & Tapanainen, 1994). Fortunately, there exist tokenization libraries for most programming languages, such as NLTK for Python and TextAnalysis for Julia, that are able to tokenize standard prose with reasonable accuracy. Unfortunately, these libraries perform less well for the text in 10-Ks and 10-Qs due to the high complexity of the text (e.g. many abbreviations, enumerations, company/product names, websites). Therefore, we develop a tokenization procedure based on regular expressions that is specifically tailored to the text in 10-Ks and 10-Qs. In a random selection of reports we compare the output from both methods and find that the regular expression based method is able to tokenize the text with higher accuracy.

5.1.2 Negation

Texts often contain negators that flip the meaning of (part of) a sentence. Ignoring these negators could lead to incorrect conclusions, since managers commonly frame negative news using positive words, e.g. "did not improve" or "no new opportunities". On the other hand, managers rarely frame positive news as non negative, e.g. "not downgraded". Therefore, like Loughran and McDonald (2011), we handle occurrences of six negators (*no, not, none, neither, never, nobody*) by prepending 'NEG_' to positive words (See Section 5.3.2 for the definition of positive) in the same sentence (detected with a sentence variant of the tokenization procedure) and in a window of length 3 on both sides of the negator. In this way, we can easily distinguish between positive and negative uses of the same word.

5.1.3 Stopwords

According to Zipf's law (Powers, 1998), the frequency of any word in a document is inversely proportional to its rank in a frequency table. Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the

a bi-directional encoder developed by Google, which recently achieved state-of-the-art performance in 11 language tasks (Devlin, Chang, Lee, & Toutanova, 2018).

third most frequent word, etc. Consequently, a small amount of words already accounts for a considerable part of the total document size. Most of the very frequent words, however, do not contain much informational value, e.g. *an, him, it, to, where.* These 'stopwords' are therefore often removed to decrease complexity and dimensionality of the data. To filter out the stopwords, we use the popular stopword list of Loughran and McDonald[16], which contains more than 150 prepositions, articles, pro-nouns, etc.

5.2 Bag-of-words

After preprocessing the text in the filings we convert it into a numerical representation with the Bag-of-Words (BoW) model. BoW is a technique from the natural language processing (NLP) literature that reduces the complexity of text data by removing information about word order and context.[17] A corpus (collection of texts) is converted to BoW representation as follows. Let $\mathcal{D} = \{\boldsymbol{d}_i\}_{i=1}^N$ be a corpus of size N. Then, after preprocessing, document i can be represented as $\boldsymbol{d}_i = [w_1, w_2, \ldots, w_{n_i}]$, where w_j is the jth token of document i, and n_i is the total number of tokens in document i. The union of all (unique) tokens in all documents in the corpus is $u = \bigcup_{i=1}^N \boldsymbol{d_i}$. The term frequency of a term $t \in u$ in document i is then computed as

$$tf(\boldsymbol{d_i}, t) = \sum_{i=1}^{N_i} \mathbf{1}_{\{w_i = t\}}, \tag{16}$$

where $\mathbf{1}_{\{A\}}$ is an indicator function that is equal to 1 if the event A is true, and 0 otherwise. Finally, the document term matrix (DTM) is the term frequency of all terms in the corpus for each document stacked where the columns are the tokens and the rows are the documents. So, the ijth entry of the DTM corresponds to the ith row and the jth column, which contains the term frequency of the jth token in the ith document.

5.2.1 Example

As an example, let the corpus $\mathcal{D}$ contain the following documents:

$$A : \text{“Text analysis, although difficult, is very useful!”}$$
$$B : \text{“Text analysis is very useful!”}$$
$$C : \text{“Text analysis is difficult.”}$$

Then, preprocess the documents

$$A = [\text{text, analysis, although, difficult, very, useful}],$$
$$B = [\text{text, analysis, very, useful}],$$

[16] `https://sraf.nd.edu/textual-analysis/resources/#StopWords`

[17] Alternatively, it is possible to use a shingles or n-gram representation, which splits the text in groups of n adjacent tokens. In this way, more information is preserved about the context that words appear in, however, it also increases the dimensionality exponentially with n.

$$C = [\text{text, analysis, difficult}],$$

compute the union of all tokens

$$u = A \cup B \cup C = [\text{text, analysis, although, difficult, very, useful}],$$

compute the term frequency vectors

$$\begin{aligned} tf(A) &= [1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1], \\ tf(B) &= [1\ \ 1\ \ 0\ \ 0\ \ 1\ \ 1], \\ tf(C) &= [1\ \ 1\ \ 0\ \ 1\ \ 0\ \ 0], \end{aligned}$$

and finally construct the DTM as

$$\text{DTM} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

5.3 Variable construction

Based on the DTM, we compute i) readability, ii) sentiment, and iii) similarity variables.

5.3.1 Readability

As recommended by Bonsall et al. (2017), we use the length of the text in a filing as a simple but reasonable proxy for the filing's level of readability. The length (LEN) of document i is computed as,

$$LEN_i = \sum_{j=1}^{|u|} DTM_{i,j}, \tag{17}$$

where $|\cdot|$ is the cardinality (number of elements) of a set.

5.3.2 Sentiment

We construct tone or sentiment variables by counting words in the positive and negative financial word lists of Loughran and McDonald (2011), in short the LM wordlists, which are nowadays widely used to classify the text in 10-Ks and 10-Qs (Kearney & Liu, 2014).[18] Table 3 shows examples of words that occupy these word lists. The positive word list for example contains the words *achieved, exceptional, outperformed*, while the negative word list contains words such as *complaint, devalue, unprofitable.*

The sentiment level of document i is computed as

$$SENT_i^{count} = \sum_{j=1}^{|u|} \mathbf{1}_{\{u_j \in LM_{SENT}\}} DTM_{i,j}, \tag{18}$$

[18] https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary

Table 3: Loughran and McDonald financial word lists

Category	#	Examples
Positive	354	achieved, compliments, exceptional, leading, lucrative, outperformed, surpass
Negative	2709	abruptly, bail, breach, complaint, devalue, illiquid, overcharged, setback, unprofitable

This table presents excerpts from the positive and negative wordlists of the LM master dictionary.

where SENT = {POS, NEG}, and $\mathbf{1}_{\{u_j \in LM_{SENT}\}}$ is an indicator that is equal to 1 if the word corresponding to the jth column of the DTM is in the respective word list and 0 otherwise. In this way, texts that contain more positive (negative) words are considered to be more positive (negative). Counts, however, might not be the best way to estimate the sentiment level of a text, since longer texts by construction contain more words. Consequently, longer texts will have higher sentiment word counts even though they are not necessarily more positive or negative. Hence, we correct for document length by dividing the sentiment variables by the length of the document, such that the variables can be interpreted as the percentage of positive or negative words

$$SENT_i = \frac{SENT_i^{count}}{LEN_i}. \tag{19}$$

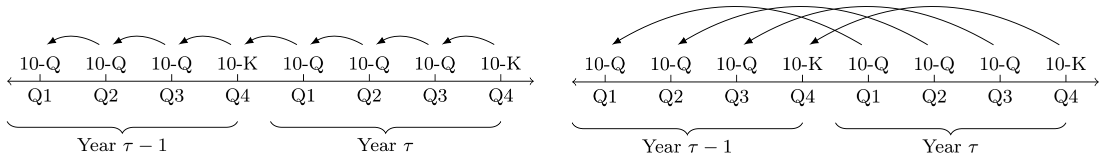
Furthermore, we also compute the net sentiment or polarity of a text, defined as

$$POL = POS - NEG, \tag{20}$$

to assess the overall sentiment of a text. The sentiment variables POS and NEG are percentages that are bounded between 0 and 1. Consequently, the polarity of a text ranges between -1 and 1, where POL = 0 if the text contains an equal amount of positive and negative words, POL $>$ 0 if the text contains more positive than negative words, and POL $<$ 0 if the text contains more negative than positive words. The negative word list contains almost eight times as many words as the positive word list (2709 vs. 354 words). Therefore, by construction, negative word counts are inflated relative to positive word counts, and hence, the polarity variable will be biased downward.

5.3.3 Changes in readability and sentiment

Next to the levels of readability and sentiment, we also compute filing-over-filing changes in the levels of readability and sentiment. This means, to compare the level of readability or sentiment of a filing with the corresponding values in a previous filing of the same company. There are two viable options to compute filing-over-filing changes: either as quarterly or yearly changes (see Figure 4). Quarterly changes allow the use of the most up-to-date information that is available. However, then variables computed from 10-Ks and 10-Qs would have to be compared, which does not make sense given their fundamental differences. Therefore, similar to Cohen et al. (2018), we use year-over-year changes such



(a) Quarterly changes

(b) Yearly changes

Fig 4: Changes in reports over time

that a 10-K is always compared with previous year's 10-K and a 10-Q with previous year's 10-Q in the same quarter. We denote the yearly change in a variable by Δ, such that e.g. the yearly change in positive words between the text in a 10-K or 10-Q in year $\tau - 1$ and τ is given by ΔPOS_τ.

5.3.4 Similarity

To gauge the degree of commonality between the reports of a company over time we use two similarity measures from the information retrieval literature (Singhal et al., 2001). Similar as in Section 5.3.3 a 10-K is always compared with previous year's 10-K and a 10-Q with previous year's 10-Q in the same quarter. The first measure that we use is the cosine similarity coefficient (COS), which measures the angle between two term frequency vectors $\boldsymbol{x} = \text{DTM}_i$ and $\boldsymbol{y} = \text{DTM}_j$ and is defined as

$$COS_{ij} = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} = \frac{\sum_{i=1}^{|u|} x_i y_i}{\sqrt{\left(\sum_{i=1}^{|u|} x_i^2\right)}\sqrt{\left(\sum_{i=1}^{|u|} y_i^2\right)}}, \tag{21}$$

where $\|\cdot\|$ is the Euclidean (ℓ_2) norm. The cosine similarity can be interpreted as a score for the degree of shared words between two texts that can range from 0 to 1. Two documents that are identical receive a 1, two documents with no common words receive a 0, and other pairs of documents are in between. The second measure is the Jaccard similarity coefficient (JAC), which for unique word vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of documents i and j is defined as,

$$JAC_{ij} = \frac{|\boldsymbol{u} \cap \boldsymbol{v}|}{|\boldsymbol{u} \cup \boldsymbol{v}|}, \tag{22}$$

where $\cap$ is the set intersection operator, and $\cup$ is the set union operator. The Jaccard similarity can be interpreted as a score for the degree of shared unique words between two documents. This measure also ranges from 0 to 1, where 1 is identical and 0 completely different. Alternatively, both these similarity measures can be thought of as a distance measure by inverting the scores.

5.4 Descriptive statistics

Companies file their 10-Ks and 10-Qs continuously throughout the year, month, and day. Hence, the filing data is of a much higher frequency than the equity and corporate bond data. Therefore, we convert all text variables to a monthly frequency. To avoid any look-ahead bias, we only include the text variables in the sample in the month after publication, such that, for example, text variables computed from the text in a filing published at the start of May can only be used from June onwards.

Descriptive statistics of the text variables computed from text in 10-Ks and 10-Qs over the sample period 1994 to 2017 are shown in Table 4. The average filing in the combined equity, HY, and IG sample contains a total of 17,131 words. The 5% smallest reports contain less than 2,236 words, while the longest reports contain more than 49,166 words. The large dispersion in size can partially be explained by time-variation, since at the start of the sample, the average filing contains roughly 5,000 words, whereas at the end it contains more than five times as many words. The five fold increase in length does not necessarily mean that filings have become five times more informative however. The SEC (2013) reports that boilerplate text accounts for a substantial portion of the increase. Additionally, academic literature suggests that some of the increase in length is attributable to managers purposefully seeking to obfuscate information in filings, possibly to mask poor performance (Li, 2008; Rogers, Van Buskirk, & Zechman, 2011).

Table 4: Descriptive statistics for text variables from the complete text in 10-Ks and 10-Qs.

		Quantiles					Subperiods				
	Mean	0.05	0.25	0.50	0.75	0.95	‘94-‘96	‘97-‘02	‘03-‘07	‘08-‘12	‘13-‘17
I. Readability											
LEN	17,131	2,236	6,279	12,334	21,910	49,166	4,936	8,495	18,013	23,903	25,913
ΔLEN	2,075	-3,544	-397	649	2,350	11,988	900	1,645	2,962	2,234	1,702
II. Sentiment (%)											
POL	-1.36	-0.36	-0.83	-1.25	-1.79	-2.76	-0.81	-1.15	-1.50	-1.60	-1.52
POS	0.98	0.47	0.74	0.94	1.17	1.59	0.87	0.92	1.00	1.03	1.04
NEG	2.34	0.83	1.57	2.20	2.96	4.35	1.67	2.06	2.50	2.62	2.55
ΔPOL	-0.06	0.54	0.13	-0.02	-0.22	-0.79	-0.09	-0.15	0.02	-0.08	-0.02
ΔPOS	0.04	-0.33	-0.09	0.01	0.13	0.54	0.04	0.04	0.06	0.02	0.03
ΔNEG	0.10	-0.87	-0.22	0.04	0.35	1.33	0.14	0.19	0.04	0.10	0.05
III. Similarity											
COS	0.897	0.548	0.887	0.944	0.974	0.991	0.822	0.856	0.888	0.927	0.943
JAC	0.680	0.283	0.603	0.713	0.803	0.891	0.565	0.598	0.659	0.733	0.779

This table contains descriptive statistics (grand mean, distribution, and mean over time) for text variables computed from the text in 10-Ks and 10-Qs for companies in the combined MSCI and BMI indexes, and Barclays Investment Grade and High Yield Corporate Bond indexes. The statistics are first computed in the cross-section and then averaged over the time-series (1994M1-2017M12).

Filings on average have a polarity of -1.36% since they contain 0.98% positive words and 2.34% negative words. There is quite some dispersion in sentiment scores between reports. The 5% least positive (negative) reports contain 0.47% (0.83%) positive (negative) words, and the 5% most positive (negative) reports contain 1.59% (4.35%) positive (negative) words. Moreover, only five percent of the reports have more positive than negative words, which is as expected given the difference in word list size. Over time, the average polarity has decreased by 0.71% since negative sentiment has increased with 0.88% while positive sentiment has only increased with 0.17%.

The cosine and Jaccard similarity coefficients are on average 90% and 68%, respectively. These reasonably high similarity scores indicate that the text in 10-Ks and 10-Qs contain a lot of stale information. The relatively low level of dispersion in similarity coefficients shows that in most cases reports are very similar to preceding reports. Over time, the reports also have become increasingly more similar. At the start of the sample, the average cosine (Jaccard) similarity is 82% (57%), whereas at the end it is 94% (78%). This increase in similarity is a logical consequence of the increase in length, driven by an increase in stale information. The overall difference between cosine and Jaccard similarity coefficients is considerable. The values for Jaccard similarity are always lower which makes sense given that the impact of addition or removal of a word is much higher for Jaccard similarity.

Table 5 shows the Spearman's rank correlation coefficients between the text variables. Overall, the readability, sentiment, and similarity variables have low correlations. The absence of high correlations between the (changes in) readability, sentiment, and similarity variables indicates that they actually capture independent aspects of the text in 10-Ks and 10-Qs. Only within the groups of variables there are higher correlations. The total filing length has a correlation of 42% with changes in total filing length, polarity has correlations of 51% and -65% with positive and negative sentiment respectively, the change in polarity is almost completely driven by the change in negative sentiment (correlation of 93%), and lastly, the similarity variables have a correlation of 73%.

Table 5: Correlation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) LEN	1.00									
(2) ΔLEN	0.47	1.00								
(3) POL	-0.08	-0.02	1.00							
(4) POS	0.30	0.12	0.51	1.00						
(5) NEG	0.34	0.14	-0.65	0.23	1.00					
(6) ΔPOL	-0.10	-0.22	0.28	0.06	-0.28	1.00				
(7) ΔPOS	0.20	0.29	0.21	0.36	0.09	0.18	1.00			
(8) ΔNEG	0.18	0.34	-0.19	0.07	0.32	-0.93	0.22	1.00		
(9) COS	0.31	-0.03	0.08	0.16	0.08	0.00	-0.04	-0.02	1.00	
(10) JAC	0.37	-0.03	0.07	0.13	0.06	0.00	-0.03	-0.02	0.73	1.00

This table shows the Spearman rank correlations between the text variables. The correlations are first computed in the cross-section and then averaged over the time-series (1994M1-2017M12).

The descriptive statistics of the text variables computed from the complete filing, MDA, and RF in 10-Ks and 10-Qs are shown in Table 6. The 10-K variants are all considerably larger than their 10-Q counterparts. The complete text in 10-Ks is almost three times as large as in the 10-Q, while the MDA and RF are both about two times as large. The levels of sentiment in the MDA are comparable to that of the complete filing (1.14% positive and 2.06% negative). In contrast, the RF is of a more negative nature than the general tone of the complete filing (1.60% positive and 6.15% negative). The MDA and RF sections also have high cosine and Jaccard similarity scores. The only exception is the RF in 10-Qs, which has a lower similarity score.

Table 6: Descriptive statistics for text variables from text in subsections of 10-Ks and 10-Qs.

	Filing			MDA			RF		
	All	10-K	10-Q	All	10-K	10-Q	All	10-K	10-Q
I. Readability									
LEN	17,131	32,477	12,125	5,623	8,041	4,817	3,785	6,117	3,008
ΔLEN	2,075	5,223	1,043	735	1,508	472	900	1,353	569
II. Sentiment (%)									
POL	-1.36	-1.27	-1.40	-0.91	-1.00	-0.88	-4.02	-4.55	-3.64
POS	0.98	1.08	0.95	1.14	1.10	1.16	1.30	1.60	1.08
NEG	2.34	2.35	2.34	2.06	2.10	2.04	5.32	6.15	4.71
ΔPOL	-0.06	-0.10	-0.05	-0.11	-0.19	-0.08	-0.82	-0.91	-0.75
ΔPOS	0.04	0.07	0.03	0.10	0.17	0.07	0.32	0.30	0.33
ΔNEG	0.10	0.17	0.08	0.21	0.36	0.15	1.13	1.22	1.07
III. Similarity									
COS	0.897	0.895	0.899	0.879	0.917	0.866	0.756	0.949	0.692
JAC	0.680	0.724	0.669	0.578	0.612	0.566	0.516	0.670	0.464

This table contains descriptive statistics for text variables computed from the complete filing, MDA, and RF in 10-Ks and 10-Qs for companies in the combined MSCI and BMI indexes, and Barclays Investment Grade and High Yield Corporate Bond indexes. The statistics are first computed in the cross-section and then averaged over the time-series (1994M1-2017M12).

6 Empirical results

This section presents and discusses the results for the regression and investment analysis of the readability, sentiment, and similarity text variables computed from the text in 10-Ks and 10-Qs, for equities, and HY and IG corporate bonds.

6.1 Risk and return predictability

To test the risk and return predictability of text variables computed from the text in 10-Ks and 10-Qs, we perform panel regressions. Given the inherent unbalancedness of the filing dataset over the months we first aggregate all data to a quarterly level by simply taking the only observation in the quarter for the text variables and taking the quarter's end value for all other variables (notice that this does not decrease the size of the sample). Additionally, in an effort to reduce the impact of outliers we normalize all variables by converting them to robust z-scores in the cross-section and capping the respective scores at absolute values of 3 (Iglewicz & Hoaglin, 1993).[19]

[19]Robust z-scores are computed as $z_i = \frac{x_i - \text{median}_x}{1.483 \times \text{MAD}_x}$ for elements $x_1, \ldots, x_N$ of a variable x, where median_x denotes the median of x and MAD_x denotes the median absolute deviation of x. The factor 1.483

6.1.1 Risk predictability

To test the (long-term) risk predictability of the text variables, we perform regressions of future 1-year stock and bond return volatility on text variables. We carry out two variations of this regression. First, we test all text variables separately. Second, we combine a selection of readability, sentiment, and similarity variables in one regression to see if the different types of variables capture different dimensions of risk. For both variants, size, value, and momentum are included as control variables. Moreover, past 1-year stock return volatility and DTS are respectively included as additional regressors for the equity and corporate bond regressions. For bonds we use DTS instead of 1-year bond return volatility because it is a very good proxy of credit volatility and because it does not depend on data of up to a year old which allows us to keep very young bonds (age less than 1 year) in the sample. The results for the regressions are shown in Table 7. Robust t-statistics are provided in parentheses for a test on the difference from zero for the regression coefficients.

Table 7: Regression results for future 1-year volatility

	Equity		HY		IG	
	Indiv	Comb	Indiv	Comb	Indiv	Comb
LEN	**0.64**	**0.46**	0.11	-0.28	**1.78**	**2.01**
	(2.44)	**(2.20)**	(0.13)	(-0.28)	**(2.61)**	**(2.84)**
ΔLEN	**1.10**	**0.69**	**2.05**	**1.80**	**1.09**	-0.01
	(5.98)	**(3.30)**	**(3.46)**	**(2.57)**	**(2.10)**	(-0.03)
POL	**-0.98**	**-0.56**	**-3.07**	**-2.61**	**-2.5**	**-1.90**
	(-4.41)	**(-2.32)**	**(-3.98)**	**(-3.14)**	**(-3.84)**	**(-2.81)**
POS	**0.53**		**-2.22**		**-2.72**	
	(2.30)		**(-2.50)**		**(-4.16)**	
NEG	**1.46**		1.05		0.57	
	(5.22)		(1.18)		(0.78)	
ΔPOL	**-1.51**	**-1.23**	**-2.38**	-1.14	**-1.77**	**-1.12**
	(-7.67)	**(-5.70)**	**(-3.57)**	(-1.59)	**(-3.50)**	**(-2.05)**
ΔPOS	**-0.36**		0.32		-0.36	
	(-2.02)		(0.54)		(-0.75)	
ΔNEG	**1.48**		**2.50**		**1.75**	
	(7.57)		**(3.74)**		**(3.43)**	
COS	**-1.15**		**-2.06**		**-1.79**	
	(-5.01)		**(-3.00)**		**(-3.05)**	
JAC	**-0.59**	**-0.45**	**-2.39**	**-1.92**	**-2.42**	**-2.72**
	(-2.55)	**(-1.99)**	**(-3.68)**	**(-2.76)**	**(-4.23)**	**(-4.63)**

This table contains regression results for panel regressions of future 1-year volatility on text variables and control variables. In the first regression (Indiv), only a single text variable is included, whereas in the second regression (Comb), a selection of text variables is included. For both variants, size, value, and momentum are included as control variables. Morover, past 1-year volatility and DTS are respectively included as additional regressors for the equity and corporate bond regressions. Robust $t-$statistics are provided in parentheses for a test on the difference from zero for the regression coefficients. Bold values indicate statistical significance at the 95% confidence level.

is the ratio of the standard deviation of a normal random variable to its MAD.

6.1.1.1 Equities

The first column presents the results for the equity volatility regressions with a single text variable. Total filing length is positively related to future 1-year volatility, with a significant coefficient of 0.64 (2.44). This result is in line with the idea that longer 10-Ks and 10-Qs are more complex disclosures that contain more information to process leading to higher future stock return volatility (Li, 2008). The year-over-year change in total filing length is also positively related to future volatility, with a significant coefficient of 1.10 (5.98). This indicates that incremental information adds complexity which leads to higher future volatility.

The polarity of the text in 10-Ks and 10-Qs negatively affects future volatility, with a significant coefficient of -0.98 (-4.41). To better understand the driving force(s) behind this polarity effect, we decompose it into separate positive and negative components. Positive and negative sentiment both have positive effects on future volatility, with significant coefficients of 0.53 (2.23) and 1.46 (5.22) respectively. This shows that both positive and negative words increase uncertainty. Hence, the polarity effect can be attributed to the difference in strength of the relation between positive and negative words and future volatility. The volatility-word tone linkage found here is consistent with the findings of Loughran and McDonald (2011). Changes in polarity also negatively affect future volatility, with a significant coefficient of -1.51 (-7.67). Again, this effect can be decomposed into separate positive and negative parts. Changes in positive sentiment are negatively related to future volatility, with a significant coefficient of -0.36 (-2.02), while changes in negative sentiment have a positive effect, with a significant coefficient of 1.48 (7.57). Hence, the impact of incremental information on future volatility depends on the sentimental content of the changes.

Cosine and Jaccard similarity are both negatively related to future volatility, with significant coefficients of -1.15 (-5.01) and -0.59 (-2.55) respectively. Therefore, companies with many changes in their 10-Ks and 10-Qs have higher volatility than companies without many changes. In conjunction with the strong results for changes in readability and sentiment variables this suggests that especially the content of the changes is highly informative for future volatility. This makes sense given that over time stale information should gradually get incorporated into the stock price.

The second column presents the results for the equity volatility regression with a combination of readability, sentiment, and similarity variables. In this regression, we find that all text variables stay (highly) significant, which indicates that the different types of text variables are not proxies of the same underlying theme, but that they actually capture distinct features of the text in the filings.

Altogether, these results confirm the results from previous studies on the predictive power of readability, sentiment, and similarity variables for future stock price volatility.

6.1.1.2 Corporate bonds

Next, we examine whether the text variables are also associated with future 1-year corporate bond return volatility. If investors truly underreact to the information contained in 10-Ks and 10-Qs then similar results as for equities are expected for the readability and similarity variables. For the sentiment variables the results might differ however because of the fundamental differences between the markets: bondholders are generally more concerned with downside risk, credit risk, and a company's ability to repay debt (conservative view), than with growth (upside oriented), which is the main objective of shareholders.

The third and fifth column respectively present the results for the HY and IG corporate bond volatility regressions with a single text variable. Total filing length is positively related to future volatility. In HY, with an insignificant coefficient of 0.11 (0.13), while in IG, with a significant coefficient of 1.78 (2.61). The change in total filing length also positively affects future volatility, with significant coefficients of 2.05 (3.46) in HY and 1.09 (2.10) in IG. The relation between the readability variables and future volatility thus has the same sign (positive) as in the equity case. This shows that share- and bondholders of a company react similarly to (changes in) the readability of the text in 10-Ks and 10-Qs.

Polarity is negatively related to future volatility, with significant coefficients of -3.07 (-3.98) in HY and -2.50 (-3.84) in IG. Changes in polarity are also negatively related to future volatility, with significant coefficients of -2.38 (-3.57) in HY and -1.77 (-3.50) in IG. The relation between (changes in) polarity and future volatility thus has the same sign (negative) as in the equity case. Interestingly, the impact of positive and negative words is not the same as for equities. The fraction of positive words has a negative effect on future volatility, with significant coefficients of -2.38 (-3.57) in HY and -1.77 (-3.50) IG, while the fraction of negative words is not significantly related to future volatility. Moreover, changes in the fraction of negative words have a positive effect on future volatility, with significant coefficients of -2.38 (-3.57) in HY and -1.77 (-3.50) IG, while changes in the fraction of positive words are not significantly related to future volatility.

Cosine and Jaccard similarity both negatively affect future volatility. This is for cosine similarity with significant coefficients of -2.06 (-3.00) in HY and -1.79 (-3.05) in IG, while for Jaccard similarity, with significant coefficients of -2.39 (-3.68) in HY and -2.42 (-4.23) in IG. The relation between the similarity variables and future volatility thus has the same sign (negative) as in the equity case. This shows that share- and bondholders of a company react similarly to changes in the text in 10-Ks and 10-Qs.

The fourth and sixth column present the results for the HY and IG corporate bond volatility regressions with a combination of readability, sentiment, and similarity variables. In these regressions, we find that (almost) all text variables that were significant in the individual regressions stay significant, which is again a sign that the text variables capture distinct features of the text.

Altogether, the consistent results for risk predictability over the three different asset classes provide strong evidence that the text in 10-Ks and 10-Qs actually contains material information for capital markets and that previously documented findings for the stock market are not simply the result of data mining (Harvey et al., 2016).

6.1.2 Return predictability

To test the return predictability of the text variables, we perform regressions of future 3-month stock and corporate bond return (return until next filing deadline) on text variables. We perform two variations of this regression in the same manner as in Section 6.1.1. The only difference is that now beta and DTS-beta are respectively included as additional regressors for the equity and corporate bond regressions. For bonds we use DTS-beta instead of a bond return based beta because it is a very good proxy of market risk and because it does not depend on data of up to three years old which allows us to keep young bonds (age less than 3 year) in the sample. The results for the regressions are shown in Table 8. Robust t-statistics are provided in parentheses for a test on the difference from zero for the regression coefficients.

6.1.2.1 Equities

The first column presents the results for the equity return regressions with a single text variable. The results show that there is no significant relation between the readability variables and returns. Moreover, they also show that there is no significant relation between the sentiment variables and returns. This is consistent with the literature, since the readability variables have mostly been linked to volatility (not return) and there is only evidence of sentiment variables predicting returns in a short window after the file date of a 10-K or 10-Q (Loughran & McDonald, 2011). In contrast, we find that the similarity variables are positively related to future return. Cosine and Jaccard similarity have significant coefficients of 0.21 (2.30) and 0.25 (3.13) respectively. So, companies that do not make large adjustments to the text in their 10-Ks and 10-Qs not only have lower stock price volatility but also higher future stock returns than companies that do make many adjustments, regardless of the content of the changes. This result is in line with the findings of Cohen et al. (2018) who claim that changes to the language and construction of 10-Ks and 10-Qs have strong (negative) implications for a company's future stock returns.

Table 8: Regression results for future 3-month return

	Equity		HY		IG	
	Indiv	Comb	Indiv	Comb	Indiv	Comb
LEN	0.12	0.11	**0.39**	**0.45**	0.03	0.03
	(0.93)	(0.83)	**(2.63)**	**(2.92)**	(0.63)	(0.59)
ΔLEN	-0.12	-0.12	-0.01	-0.12	0.00	0.01
	(-1.21)	(-1.2)	(-0.07)	(-1.04)	(-0.01)	(0.09)
POL	-0.02	-0.01	-0.02	-0.02	0.01	0.01
	(-0.21)	(-0.10)	(-0.18)	(-0.16)	(0.36)	(0.14)
POS	0.18		0.09		0.02	
	(1.60)		(0.77)		(0.72)	
NEG	0.16		-0.11		-0.02	
	(1.12)		(0.81)		(-0.46)	
ΔPOL	0.03	0.01	0.07	0.08	0.04	0.04
	(0.29)	(0.09)	(0.44)	(0.50)	(1.06)	(1.07)
ΔPOS	-0.05		-0.04		0.03	
	(-0.55)		(-0.30)		(1.00)	
ΔNEG	-0.06		-0.10		-0.03	
	(-0.59)		(-0.65)		(-0.78)	
COS	**0.21**		0.27		**0.09**	
	(2.30)		(1.87)		**(2.54)**	
JAC	**0.25**	**0.21**	0.18	0.15	**0.08**	**0.09**
	(3.13)	**(2.89)**	(1.35)	(1.2)	**(2.39)**	**(2.47)**

This table contains regression results for panel regressions of future 3-month return on text variables and control variables. In the first regression (Indiv), only a single text variable is included, whereas in the second regression (Comb), a selection of text variables is included. For both variants, size, value, and momentum are included as control variables. Morover, beta and DTS-beta are respectively included as additional regressors for the equity and corporate bond regressions. Robust $t-$statistics are provided in parentheses for a test on the difference from zero for the regression coefficients. Bold values indicate statistical significance at the 95% confidence level.

The second column presents the results for the regression with a combination of readability, sentiment, and similarity variables. In this regression, we find that the similarity variable stays significant when readability and sentiment variables are included as additional control variables.

Altogether, these results confirm the results from previous studies on the predictive power of readability, sentiment, and similarity variables on future stock return.

6.1.2.2 Corporate bonds

Next, we examine whether the text variables are also associated with future 3-month corporate bond return. The third and fifth column respectively present the results for the HY and IG corporate bond return regressions with a single text variable. The results show that similar as for equities there is no significant relation between the readability or sentiment variables and returns. With the only exception of the total filing length in HY, which has a significant coefficient of 0.39 (2.63). Conversely, we find that the similarity variables are positively related to future corporate bond return. For HY, the relation is modest but insignificant with coefficients of 0.27 (1.87) and 0.18 (1.35) for cosine and Jaccard similarity respectively, while for IG, the relation is fairly strong with significant coefficients of 0.08 (2.39) and 0.09 (2.47). Therefore, companies that do not make large adjustments to the text in their 10-Ks and 10-Qs not only have lower corporate bond volatility but also higher future corporate bond returns than companies that do make many adjustments, regardless of the content of the changes.

The fourth and sixth column present the results for the regressions with a combination of readability, sentiment, and similarity variables. In these regressions, we find that the similarity variables that were significant in the individual regressions stay significant when readability and sentiment variables are included as additional control variables.

Altogether, the consistent results for return predictability over the three different asset classes even further strengthen the belief that the text in 10-Ks and 10-Qs actually contains material information for capital markets.

6.2 Characteristics sorted portfolios

To further test the relation between the text variables and risk and return, we perform backtests for characteristics sorted portfolios based on the text variables. Remember that the text variables have four data points per company per year, since only every quarter a new 10-K or 10-Q is filed by a company. Therefore, for each variable, we fill the months until the next publication with the last value, up to a maximum of four months. In this way, stocks and bonds can be included in a portfolio at all times instead of only 4 times per year (the months after filing a new report). The results for backtests of the readability, sentiment, and similarity strategies with a one month holding period over the historical sample from 1997 to 2017 are shown in Table 9 for equally-weighted quintile long-only portfolios and a hedging portfolio. For each strategy, annualized mean return, volatility, and Sharpe ratio are reported, together with robust t-statistics in parenthesis from a difference test against the respective market statistics.

6.2.1 Readability variables

Panel A presents the results for the strategies based on readability variables. There are no portfolios based on readability variables that significantly out- or underperform the market in any of the three asset classes, except for one of them. The bottom portfolio for total filing length in HY has a rather high return of 5.84% (2.53). The lack of a (monotonic) pattern for the returns of the other portfolios (Q1-Q4) in HY and the absence of outperforming portfolios in the other asset classes indicates that companies with better readable 10-Ks and 10-Qs generally do not generate significantly better returns. In contrast, for the

Table 9: Return statistics

		Equity						HY						IG					
Panel A: Readability																			
		Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
LEN	Mean return	8.61%	9.29%	9.11%	8.76%	9.73%	-1.12%	2.06%	3.03%	2.82%	2.56%	**5.84%**	-3.79%	0.6%	0.77%	1.0%	0.84%	0.94%	-0.34%
		(-0.61)	(0.28)	(0.01)	(-0.68)	(0.46)	–	(-1.88)	(-0.33)	(-0.82)	(-0.9)	**(2.53)**	–	(-0.85)	(-0.29)	(1.17)	(0.05)	(0.27)	–
	Volatility	**16.75%**	**17.39%**	18.39%	19.01%	**20.7%**	8.59%	**9.28%**	9.83%	10.25%	**11.32%**	**12.6%**	5.6%	**3.88%**	**4.01%**	4.46%	4.99%	**5.67%**	2.58%
		(-3.72)	**(-2.95)**	(0.98)	(1.47)	**(3.74)**	–	**(-3.54)**	(-1.59)	(-0.11)	**(2.88)**	**(3.16)**	–	**(-3.01)**	**(-2.30)**	(0.07)	(1.59)	**(2.05)**	–
	Sharpe ratio	0.51	0.53	0.5	0.46	0.47	-0.13	0.22	0.31	0.27	0.23	**0.46**	-0.68	0.15	0.19	0.22	0.17	0.17	-0.13
		(0.31)	(0.95)	(-0.21)	(-1.56)	(-0.56)	–	(-1.66)	(-0.15)	(-0.82)	(-1.62)	**(2.57)**	–	(-0.7)	(0.16)	(1.03)	(-0.43)	(-0.23)	–
ΔLEN	Mean return	9.26%	9.52%	10.01%	9.44%	8.5%	0.76%	3.74%	3.41%	3.47%	3.79%	4.01%	-0.27%	0.75%	0.82%	1.06%	0.71%	0.95%	-0.2%
		(-0.15)	(0.25)	(1.74)	(0.17)	(-0.88)	–	(0.09)	(-0.42)	(-0.36)	(0.17)	(0.29)	–	(-0.68)	(-0.17)	(0.85)	(-0.98)	(0.2)	–
	Volatility	18.52%	**17.29%**	**17.35%**	17.95%	**19.96%**	4.97%	10.1%	**9.15%**	**9.6%**	10.51%	**13.04%**	5.92%	4.54%	4.02%	**3.93%**	4.54%	**5.53%**	2.16%
		(1.7)	**(-3.35)**	**(-3.72)**	(-0.37)	**(3.43)**	–	(-0.25)	**(-3.72)**	**(-2.35)**	(0.86)	**(3.72)**	–	(0.52)	(-1.86)	**(-3.72)**	(0.65)	**(2.27)**	–
	Sharpe ratio	0.5	0.55	**0.58**	0.53	**0.43**	0.15	0.37	0.37	0.36	0.36	0.31	-0.05	0.17	0.2	**0.27**	0.16	0.17	-0.09
		(-0.63)	(1.0)	**(2.28)**	(0.27)	**(-2.31)**	–	(0.15)	(0.23)	(-0.01)	(-0.03)	(-0.91)	–	(-0.68)	(0.28)	**(2.48)**	(-1.08)	(-0.44)	–
Panel B: Sentiment																			
		Equity						HY						IG					
		Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
POL	Mean return	9.48%	9.45%	9.08%	8.75%	8.84%	0.64%	3.07%	2.1%	3.64%	3.84%	3.66%	-0.59%	0.92%	0.95%	0.87%	0.72%	0.7%	0.22%
		(0.53)	(0.64)	(-0.08)	(-0.54)	(-0.38)	–	(-0.25)	(-1.8)	(0.62)	(0.86)	(0.44)	–	(0.53)	(0.73)	(0.19)	(-0.72)	(-0.43)	–
	Volatility	**16.89%**	**17.45%**	18.59%	**19.59%**	**19.32%**	5.83%	**8.65%**	**9.32%**	**10.99%**	**11.62%**	**12.49%**	5.92%	**4.08%**	**4.14%**	4.53%	4.53%	**5.32%**	1.91%
		(-3.29)	**(-2.38)**	(0.36)	**(2.41)**	**(2.25)**	–	**(-3.72)**	**(-2.49)**	**(2.03)**	**(2.43)**	**(3.35)**	–	**(-3.43)**	**(-2.49)**	(0.28)	(0.53)	**(3.54)**	–
	Sharpe ratio	0.56	0.54	0.49	0.45	0.46	0.11	0.36	0.23	0.33	0.33	0.29	-0.1	0.23	0.23	0.19	0.16	0.13	0.11
		(1.53)	(1.26)	(-0.51)	(-1.94)	(-1.19)	–	(0.74)	(-1.76)	(0.29)	(0.27)	(-0.39)	–	(1.03)	(1.21)	(0.13)	(-0.67)	(-1.29)	–
POS	Mean return	9.89%	9.51%	9.47%	8.49%	8.15%	1.74%	3.64%	3.54%	3.11%	3.73%	2.33%	1.31%	1.14%	0.8%	0.68%	0.76%	0.79%	0.36%
		(0.58)	(0.8)	(0.59)	(-0.81)	(-1.49)	–	(0.46)	(0.53)	(-0.27)	(0.79)	(-1.44)	–	(1.49)	(-0.21)	(-1.14)	(-0.41)	(-0.27)	–
	Volatility	**21.38%**	19.01%	**17.36%**	**17.35%**	**17.46%**	8.95%	10.03%	10.02%	10.7%	10.82%	**11.69%**	5.02%	4.25%	4.62%	4.56%	4.6%	4.61%	1.48%
		(3.29)	(1.43)	**(-3.06)**	**(-2.93)**	**(-2.58)**	–	(-0.25)	(-0.88)	(0.95)	(1.24)	**(2.71)**	–	(-0.58)	(0.55)	(1.38)	(0.11)	(0.29)	–
	Sharpe ratio	0.46	0.5	0.55	0.49	0.47	0.19	0.36	0.35	0.29	0.34	**0.2**	0.26	0.27	0.17	0.15	0.17	0.17	0.24
		(-0.63)	(-0.03)	(1.3)	(-0.28)	(-0.92)	–	(0.72)	(0.75)	(-0.48)	(0.48)	**(-2.12)**	–	(1.92)	(-0.37)	(-1.24)	(-0.47)	(-0.4)	–
NEG	Mean return	8.63%	8.58%	9.17%	9.63%	9.49%	-0.85%	3.05%	**1.81%**	3.56%	4.38%	3.52%	-0.47%	0.84%	1.02%	0.85%	0.68%	0.77%	0.08%
		(-0.52)	(-0.73)	(0.1)	(0.91)	(0.24)	–	(-0.27)	**(-2.45)**	(0.54)	(1.9)	(0.25)	–	(0.07)	(1.22)	(0.07)	(-0.79)	(-0.21)	–
	Volatility	**16.73%**	**16.74%**	**17.51%**	**19.22%**	**22.47%**	10.58%	**8.96%**	9.91%	10.7%	10.97%	**12.63%**	6.18%	4.33%	4.34%	**4.27%**	4.33%	**5.36%**	1.84%
		(-2.69)	**(-2.72)**	**(-2.04)**	**(1.98)**	**(3.72)**	–	**(-3.43)**	(-1.11)	(1.27)	(1.61)	**(3.72)**	–	(-1.82)	(-1.62)	**(-1.97)**	(-0.52)	**(3.72)**	–
	Sharpe ratio	0.52	0.51	0.52	0.5	0.42	-0.08	0.34	**0.18**	0.33	0.4	0.28	-0.08	0.19	0.24	0.2	0.16	0.14	0.04
		(0.3)	(0.31)	(0.66)	(-0.02)	(-1.27)	–	(0.44)	**(-2.53)**	(0.31)	(1.65)	(-0.61)	–	(0.2)	(1.48)	(0.3)	(-0.74)	(-1.0)	–
ΔPOL	Mean return	8.94%	**10.61%**	**8.42%**	9.68%	9.1%	-0.16%	3.43%	4.16%	3.41%	3.1%	4.31%	-0.88%	1.09%	0.8%	0.96%	1.11%	0.33%	0.76%
		(-0.8)	**(2.58)**	**(-2.23)**	(0.59)	(-0.5)	–	(-0.32)	(0.75)	(-0.47)	(-0.91)	(0.6)	–	(0.99)	(-0.34)	(0.65)	(1.29)	(-1.94)	–
	Volatility	**17.33%**	17.87%	18.22%	18.59%	**18.74%**	3.29%	9.48%	9.6%	9.58%	11.01%	**12.87%**	5.96%	4.12%	4.41%	4.42%	4.69%	4.91%	1.69%
		(-2.85)	(-1.09)	(1.05)	(1.86)	**(2.35)**	–	(-1.22)	(-1.51)	(-1.66)	(1.61)	**(3.24)**	–	(-1.7)	(-0.2)	(-0.27)	(1.09)	(1.54)	–
	Sharpe ratio	0.52	**0.59**	**0.46**	0.52	0.49	-0.05	0.36	0.43	0.36	0.28	0.33	-0.15	0.26	0.18	0.22	0.24	**0.07**	0.45
		(-0.11)	**(2.99)**	**(-2.18)**	(0.08)	(-0.79)	–	(-0.0)	(1.3)	(-0.12)	(0.0)	(-0.46)	–	(1.89)	(-0.27)	(0.6)	(1.22)	**(-2.77)**	–

Continued on next page...

Table 9 – continued from previous page

ΔPOS	Mean return	9.72%	8.87%	9.5%	9.79%	8.84%	0.87%	3.49%	3.38%	3.98%	4.17%	3.4%	0.09%	1.16%	0.84%	0.96%	0.65%	0.69%	0.46%
		(0.61)	(-0.88)	(0.39)	(0.89)	(-0.92)	–	(-0.3)	(-0.59)	(0.5)	(0.7)	(-0.39)	–	(1.6)	(-0.13)	(0.52)	(-1.39)	(-0.94)	–
	Volatility	**18.44%**	17.83%	17.83%	17.9%	**18.77%**	3.63%	**11.4%**	10.28%	10.36%	9.65%	10.8%	4.47%	4.64%	4.54%	4.31%	4.72%	4.29%	1.42%
		(1.99)	(-1.25)	(-1.13)	(-0.49)	**(3.54)**	–	**(2.62)**	(0.28)	(0.53)	(-1.8)	(1.12)	–	(0.89)	(1.17)	(-1.14)	(1.53)	(-1.19)	–
	Sharpe ratio	0.53	0.5	0.53	0.55	0.47	0.24	0.31	0.33	0.38	0.43	0.31	0.02	0.25	0.18	0.22	0.14	0.16	0.33
		(0.3)	(-0.79)	(0.62)	(1.19)	(-1.84)	–	(-1.01)	(-0.65)	(0.43)	(1.28)	(-0.85)	–	(1.5)	(-0.26)	(0.79)	(-1.61)	(-0.81)	–
ΔNEG	Mean return	9.83%	9.4%	9.47%	9.12%	8.9%	0.93%	4.22%	3.93%	2.44%	3.55%	4.28%	-0.06%	0.96%	1.06%	0.8%	1.05%	0.41%	0.55%
		(1.02)	(0.1)	(0.31)	(-0.49)	(-0.65)	–	(0.76)	(0.34)	(-1.94)	(-0.24)	(0.56)	–	(0.47)	(1.24)	(-0.43)	(1.1)	(-1.67)	–
	Volatility	18.22%	**17.49%**	**17.57%**	18.03%	**19.71%**	5.15%	9.56%	**9.28%**	9.79%	**11.16%**	**12.8%**	5.71%	4.24%	**4.15%**	4.49%	4.71%	**4.93%**	1.68%
		(0.61)	**(-2.10)**	**(-2.05)**	(0.01)	**(3.43)**	–	(-1.7)	**(-3.43)**	(-0.79)	**(2.57)**	**(3.29)**	–	(-1.31)	**(-3.35)**	(0.39)	(1.2)	**(2.46)**	–
	Sharpe ratio	0.54	0.54	0.54	0.51	0.45	0.18	0.44	0.42	**0.25**	0.32	0.33	-0.01	0.23	**0.25**	0.18	0.22	**0.08**	0.33
		(0.69)	(0.61)	(0.82)	(-0.45)	(-1.9)	–	(1.45)	(1.07)	**(-2.02)**	(-0.82)	(-0.46)	–	(0.83)	**(2.04)**	(-0.45)	(0.84)	**(-2.52)**	–

Panel C: Similarity

		Equity						HY						IG					
		Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
COS	Mean return	**10.79%**	9.77%	8.85%	8.95%	8.39%	2.39%	4.45%	**4.99%**	3.28%	3.33%	**2.01%**	2.38%	**1.34%**	1.12%	0.91%	0.59%	**0.32%**	1.03%
		(2.01)	(0.87)	(-1.06)	(-0.79)	(-1.16)	–	(1.45)	**(2.52)**	(-0.99)	(-0.95)	**(-2.25)**	–	**(3.07)**	(1.63)	(0.41)	(-1.45)	**(-2.68)**	–
	Volatility	**17.13%**	17.94%	18.2%	**18.58%**	**19.16%**	5.29%	**9.14%**	10.25%	10.31%	10.85%	**12.31%**	5.40%	4.35%	4.42%	4.4%	4.62%	4.66%	1.27%
		(-3.04)	(-0.56)	(1.1)	**(2.3)**	**(3.72)**	–	**(-2.26)**	(0.13)	(0.17)	(1.66)	**(2.97)**	–	(-0.54)	(-0.16)	(-0.45)	(1.86)	(0.98)	–
	Sharpe ratio	**0.63**	0.54	0.49	0.48	**0.44**	0.45	**0.49**	**0.49**	0.32	0.31	**0.16**	0.44	**0.31**	0.25	0.21	0.13	**0.07**	0.81
		(3.05)	(1.0)	(-1.25)	(-1.44)	**(-2.26)**	–	**(2.60)**	**(2.53)**	(-0.70)	(-0.82)	**(-3.11)**	–	**(3.59)**	(1.9)	(0.46)	(-1.8)	**(-3.19)**	–
JAC	Mean return	**11.79%**	9.51%	9.6%	**8.06%**	**7.78%**	4.01%	4.14%	3.37%	3.51%	4.32%	3.09%	1.05%	**1.3%**	1.08%	0.87%	0.66%	**0.37%**	0.94%
		(3.15)	(0.37)	(0.58)	**(-2.66)**	**(-2.75)**	–	(0.71)	(-0.51)	(-0.28)	(0.85)	(-0.82)	–	**(2.96)**	(1.46)	(0.11)	(-1.09)	**(-2.12)**	–
	Volatility	18.45%	17.93%	17.82%	17.92%	**18.92%**	4.91%	**9.24%**	9.61%	**11.08%**	10.48%	**11.98%**	4.91%	**4.15%**	4.41%	4.36%	4.65%	**4.96%**	1.49%
		(1.26)	(-0.65)	(-1.13)	(-0.48)	**(3.54)**	–	**(-3.72)**	(-1.22)	**(2.02)**	(0.65)	**(2.99)**	–	**(-2.64)**	(-0.28)	(-0.77)	(1.83)	**(3.06)**	–
	Sharpe ratio	**0.64**	0.53	0.54	**0.45**	**0.41**	0.82	0.45	0.35	0.32	0.41	0.26	0.21	**0.31**	0.24	0.2	0.14	**0.07**	0.63
		(2.98)	(0.46)	(0.74)	**(-2.26)**	**(-3.7)**	–	(1.72)	(-0.2)	(-0.98)	(0.88)	(-1.91)	–	**(3.38)**	(1.61)	(0.2)	(-1.12)	**(-2.93)**	–

This table shows performance statistics of text factors for stocks in the combined MSCI and BMI indexes and corporate bonds in the U.S. Investment Grade and High Yield indexes over the period January 1997 - December 2017. Each month, factor portfolios take equally-weighted long positions in 20% of the stocks and bonds based on text variables such that the highest ranking bonds and stocks are in Q1, the 'next best' 20% are in Q2, and so forth up till the lowest ranking 20% which are in Q5. At time t, the text variables are computed from the text in 10-K and 10-Q filings published in the three preceding months. For LEN, the companies are ranked based on the length of the text (Q1: low, Q5: high); for ΔLEN, the companies are ranked based on the difference in length of the text in their 10-K or 10-Q filing published in the last three months compared to the same report last year (Q1: low, Q5: high); for POL, the companies are ranked based on the difference between POS and NEG (Q1: high, Q5: low), where for POS the companies are ranked based on positive sentiment in the text (Q1: high, Q5: low) and for NEG on negative sentiment (Q1: low, Q5: high); for ΔPOL, the companies are ranked based on the difference between ΔPOS and ΔNEG (Q1: high, Q5: low), where for ΔPOS the companies are ranked based on changes in positive sentiment in the text (Q1: high, Q5: low) and for ΔNEG on changes in negative sentiment in the text (Q1: low, Q5: high); for COS the companies are ranked based on the similarity of subsequent 10-Ks or 10Qs in terms of word frequency (Q1: high, Q5: low); for JAC the companies are ranked based on the similarity of subsequent 10-Ks or 10-Qs in terms of word occurence (Q1: high, Q5: low). Mean, volatility and Sharpe ratio are annualized. t-statistics are provided in parentheses for two-sided tests whether the mean, variance (Ledoit & Wolf, 2011), and Sharpe ratio (Ledoit & Wolf, 2008) of factor portfolios are different from the corresponding measures of the market portfolio. Bold values indicate statistical significance at the 95% confidence level.

volatility of the portfolios, there are significant deviations from the market. For total filing length, the volatility of the portfolios exhibits a monotonically increasing pattern, where companies with the shortest (longest) texts have significantly lower (higher) volatility than the market. For changes in total filing length, the pattern over the portfolios could best be described as an asymmetric U-shape. Companies with the largest absolute changes in the length of the text have significantly higher volatility than the market, with increases in length leading to higher volatility than decreases in length, while companies with the smallest absolute changes have significantly lower volatility. These results clearly show that the readability of the text in 10-Ks and 10-Qs is related to stock price volatility. However, the results for changes in the total filing length also indicate that not only changes in readability, but changes to anything in the text might actually drive volatility (remember that this is precisely what the similarity variables measure). Finally, as an investment strategy, the hedge portfolios based on the top minus the bottom readability portfolios are not really appropriate, since for most of them average (risk-adjusted) returns are even negative.

6.2.2 Sentiment variables

Panel B presents the results for the strategies based on sentiment variables. There are no portfolios based on sentiment variables that significantly out- or underperform the market in any of the three asset classes. In contrast, for the volatility of the portfolios, there are significant deviations from the market. For polarity, the volatility of the portfolios exhibits a monotonically increasing pattern, where companies with the highest (lowest) polarity have significantly lower (higher) volatility than the market. For equities, this effect is driven by both positive and negative sentiment, whereas for HY and IG, it is mainly driven by negative sentiment. For changes in polarity, there also is a monotonically increasing pattern, such that companies with the largest increase (decrease) in polarity have lower (higher) volatility than the market. However, the effect is not as strong as for polarity, especially in IG. For all three asset classes there is a strong effect for changes in negative sentiment while the effect for changes in positive sentiment is small. Moreover, for changes in negative sentiment, the volatility of the portfolios does not exhibit a monotonically increasing pattern. The bottom portfolio has significantly higher volatility than the market however the top portfolio does not have significantly lower volatility than the market. Instead, the volatility of Q2 is significantly lower than the market in all cases. This portfolio corresponds to the companies with only a small decrease in negative sentiment. Finally, as an investment strategy, the hedge portfolios based on the top minus the bottom sentiment portfolios are not very successful, since for most of them average (risk-adjusted) returns are low or even negative.

6.2.3 Similarity variables

Panel C presents the results for the strategies based on similarity variables. In all three asset classes there are portfolios based on similarity variables that significantly out- or underperform the market. For both cosine and Jaccard similarity there is a monotonically decreasing pattern in the mean return of the portfolios. Thus, changers (non-changers) have significantly lower (higher) return than the market. This is especially the case for equities and IG, while for HY, the pattern is somewhat weaker. In addition, for the volatility of the portfolios, there also are significant deviations from the market. For both cosine and Jaccard similarity there is a monotonically increasing pattern in the volatility of the portfolios. Thus, changers (non-changers) not only have significantly lower (higher) return but also significantly higher (lower) volatility than the market. Even though the similarity variables predict risk, the volatility spread (difference in top and bottom portfolio volatility) of the similarity strategies is smaller than for most of the other strategies. From a risk-adjusted perspective this makes the strategies very attractive, since the changes in return are not necessarily off-set by proportional changes in risk. This is reflected in the Sharpe ratios of the portfolios. For equities and IG, changers (non-changers) have significantly lower (higher) Sharpe ratio than the market. For HY, this is only the case for cosine similarity, since for Jaccard similarity the results are insignificant. Consequently, the hedge portfolios have excellent performance. For equities and IG, the hedge portfolios of cosine (Jaccard) similarity have a Sharpe ratio of 0.45 and 0.81 (0.82 and 0.63) respectively. Only, for HY, the hedge portfolios have somewhat lower Sharpe ratios (0.33 and 0.21 for cosine and Jaccard similarity respectively). This is not necessarily because the return of the hedge portfolios is low, but the volatility is just very high. In fact, the volatility of the hedge portfolios is higher than the volatility of the hedge portfolios for equities. However, even though the results in HY are not as strong as the result for equities and IG, the results point in the same direction.

6.3 Similarity investment strategy

The regressions and the backtests for the characteristics sorted portfolios both point to the similarity variables as the most promising text variables to use in an investment strategy, since they are predictive of return and risk for both equities and corporate bonds. In what follows, we will dive deeper into the similarity strategy to find out what drives it's profitability and to see whether the performance can even be improved.

6.3.1 Alpha statistics

The standalone performance of the similarity strategies is good. To test whether this performance is not simply the result of allocations to the market and/or other known factors such as size, value, and momentum, we compute CAPM and four-factor alpha's. The results are shown in Table 10 for both the cosine and Jaccard similarity strategies.

Table 10: Similarity measures alpha statistics

Panel A: Equity												
	COS						JAC					
	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
CAPM alpha	**2.03%**	0.53%	-0.53%	-0.62%	-1.41%	**3.44%**	**2.39%**	0.28%	0.42%	**-1.14%**	**-1.94%**	**4.33%**
	(3.09)	(1.1)	(-1.07)	(-1.18)	(-1.73)	**(2.8)**	**(2.88)**	(0.67)	(0.94)	**(-2.16)**	**(-3.5)**	**(3.55)**
Four-factor alpha	**1.82%**	0.48%	-0.5%	-0.47%	**-1.32%**	**3.14%**	**2.72%**	0.38%	0.3%	**-1.48%**	**-1.91%**	**4.63%**
	(2.79)	(1.04)	(-1.12)	(-0.81)	**(-2.37)**	**(3.16)**	**(3.64)**	(0.94)	(0.74)	**(-2.86)**	**(-3.53)**	**(4.06)**
Panel B: HY												
	COS						JAC					
	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
CAPM alpha	**1.40%**	**1.38%**	-0.15%	-0.09%	**-2.13%**	**3.53%**	0.89%	0.01%	-0.41%	0.66%	**-1.11%**	**2%**
	(2.83)	**(2.75)**	(-0.11)	(-0.05)	**(-3.42)**	**(3.76)**	(1.65)	(0.01)	(-0.73)	(0.85)	**(-2.01)**	**(2.3)**
Four-factor alpha	**1.24%**	0.36%	-0.10%	-0.31%	-0.98%	**2.22%**	0.29%	0.24%	-0.41%	0.12%	-0.23%	0.52%
	(1.98)	(0.38)	(-0.07)	(-0.61)	(-1.70)	**(2.54)**	(0.63)	(0.39)	(-0.71)	(0.13)	(-0.43)	(0.6)
Panel C: IG												
	COS						JAC					
	Q1	Q2	Q3	Q4	Q5	TB	Q1	Q2	Q3	Q4	Q5	TB
CAPM alpha	**0.51%**	0.28%	0.07%	-0.29%	**-0.57%**	**1.08%**	**0.51%**	0.24%	0.05%	-0.21%	**-0.58%**	**1.09%**
	(3.43)	(1.76)	(0.57)	(-1.55)	**(-2.87)**	**(3.47)**	**(3.61)**	(1.57)	(0.31)	(-1.2)	**(-2.86)**	**(3.99)**
Four-factor alpha	**0.51%**	0.15%	0.22%	-0.39%	-0.49%	**1%**	**0.57%**	0.03%	0.05%	-0.15%	-0.5%	**1.07%**
	(2.62)	(0.74)	(1.42)	(-1.49)	(-1.86)	**(2.52)**	**(2.78)**	(0.15)	(0.29)	(-0.57)	(-1.66)	**(2.55)**

This table contains alpha statistics for both of the similarity strategies (COS and JAC). It presents CAPM and four-factor alpha's for the long-only portfolios and the hedge portfolio. All values are annualized. t-statistics are provided in parentheses for two-sided tests whether the alpha's deviate from zero. Bold values indicate statistical significance at the 95% confidence level.

Panel A presents the results for equities. The cosine similarity strategy has a four-factor alpha (from here on referred to as alpha) of 1.82% (2.79) for the top portfolio, -1.32% (-2.37) for the bottom portfolio, and 3.14% (3.16) for the hedge portfolio. This means that the outperformance of the strategy is not only statistically but also economically significant. Moreover, the Jaccard similarity strategy has an alpha of 2.72% (3.64) for the top portfolio, -1.91% (-3.53) for the bottom portfolio, and 4.63% (4.06) for the hedge portfolio. The outperformance of the Jaccard similarity strategy thus also is economically significant, and even more so than for the cosine similarity strategy.

Panel B and C present the results for HY and IG. For HY, only the hedge portfolio of the cosine similarity strategy has a significant alpha of 2.22% (2.54). For IG, the cosine similarity strategy has an alpha of 0.51% (2.62) for the top portfolio and 1.00% (2.52) for the hedge portfolio, and the Jaccard similarity strategy has an alpha of 0.57% (2.78) for the top portfolio and 1.07% (2.55) for the hedge portfolio. The similarity strategies thus have economical significance for corporate bonds, especially in the long-short context.

In sum, we find that the similarity strategies generate an alpha in excess of 3% for equities, 2% for HY, and 1% for IG, even when controlled for other factors. Hence, the text in 10-Ks and 10-Qs contains material information for future stock and bond returns that is not (completely) captured by quantitative factors.

6.3.2 Spanning regressions

To test whether one of the similarity strategies performs better than the other we carry out a set of spanning regressions. The spanning regressions are of the following form

$$R_t^y = \alpha^* + \lambda R_t^x + \beta_1 RMRF_t + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 UMD_t + \varepsilon_t, \qquad (23)$$

where R_t^y denotes the return of strategy y at time t, α^* is the spanning alpha, λ is the coefficient of the return of strategy x on the return of strategy y, and $RMRF_t$, SMB_t, HML_t, and UMD_t are the returns of the stock or bond factors (see Section 3.2). The spanning alpha obtained from this regression can be interpreted as the excess alpha of strategy y beyond strategy x (and the control variables). A positive value for the α^* means that strategy y has added value over strategy x, while a non-positive α^* means that strategy y is dominated by strategy x.

Table 11 shows the results for spanning regressions of the returns of each of the similarity strategies on the other. In general, we find that the returns of one of the similarity strategies predict the returns of the other strategy. This is not surprising, since the similarity variables are closely related and also highly correlated.

Table 11: Spanning regressions similarity strategies

	Equity		HY		IG	
	COS	*JAC*	*COS*	*JAC*	*COS*	*JAC*
Spanning alpha	0.19%	**2.88%**	1.00%	-0.36%	0.65%	0.38%
	(0.16)	**(2.87)**	(1.38)	(-0.67)	(1.89)	(1.77)
COS		**0.47**		**0.73**		**0.54**
		(4.93)		**(19.16)**		**(2.46)**
JAC	**0.55**		**0.86**		**0.40**	
	(8.33)		**(7.00)**		**(2.38)**	

This table contains regression results for spanning regressions of the hedge portfolio return of either of the similarity strategies (COS and JAC) on a constant, the hedge portfolio return of the other similarity strategy and controls (size, value, and momentum). Bold values indicate statistical significance at the 95% confidence level.

For equities, the returns of the Jaccard similarity strategy explain nearly all alpha of the cosine similarity strategy, as only 0.19% (0.16) alpha remains after controlling for Jaccard similarity. The other way around, the returns of the cosine similarity strategy only explain close to half of the alpha of the Jaccard similarity strategy, as a significant alpha of 2.88% (2.87) still remains after controlling for cosine similarity. The superiority of the Jaccard similarity strategy for equities is also documented by Cohen et al. (2018).

For HY, the returns of the Jaccard similarity strategy only explain part of the alpha of the cosine similarity strategy, since still 1.00% (1.38) alpha remains after controlling for Jaccard similarity. Moreover, we previously found that the Jaccard similarity strategy has a positive four-factor-alpha in HY, however, when controlled for the returns of the cosine similarity strategy, a negative alpha of -0.36% (-0.67) remains. Here, the Jaccard similarity strategy thus performs worse than the cosine similarity strategy. For IG, the returns of the Jaccard and cosine similarity strategy both explain part of the alpha of the other, as sizeable alpha's of 0.65% (1.89) and 0.38% (1.77) remain respectively.

Overall, we find that there is no winner strategy that performs best for all asset classes. Instead, the positive alpha's that remain after controlling for the other similarity strategy show that the strategies have incremental performance over each other. Therefore, an improved strategy with better and/or more robust (out-of-sample) performance could be constructed by combining the cosine and Jaccard similarity variables (Ravazzolo, 2007).

6.3.3 Composite similarity strategy

Based on the spanning regression results, we construct a composite similarity variable (SIM) as the equally-weighted combination of the cosine and Jaccard similarity variables, where the cosine and Jaccard similarity variables have first been converted to robust z-scores. This variable is a measure of commonality between two documents that takes into account both the occurrence and frequency of words. In this way, it represents a more general concept of similarity than any specific implementation of similarity would. As before, we perform backtests for portfolios based on this variable to test it's performance.

Table 12 shows the performance statistics for a backtest of the composite similarity strategy. For equities and HY, performance of the hedge portfolio is slightly below the performance of the best similarity strategy in each asset class (the Jaccard similarity strategy for equities and cosine similarity strategy for HY), while in IG, performance is even slightly better than for both similarity strategies. These outcomes are in line with the spanning regression results in Section 6.3.2. The performance of the hedge portfolios comes from both the long and the short legs for equities and IG, while it comes mostly from the short leg for HY. Moreover, the top portfolios have low tracking-error and high outperformance, especially for equities and IG. This makes the strategy interesting for (semi-)active investors who's objective it is to outperform a given benchmark, usually the market portfolio, with only limited risk exposure.

The profitability of investment strategies based on similarity variables extracted from the text in 10-Ks and 10-Qs was already shown in the literature for equities. We now show that this performance also transfers to other asset classes, in particular the IG and HY corporate bond markets. In what follows, we take a more detailed look at the strategy to

Table 12: Performance statistics for the composite similarity strategy

Panel A: Equity	Q1	Q2	Q3	Q4	Q5	TB
I. Return statistics						
Mean return	**11.24%**	9.85%	9.3%	8.75%	**7.61%**	3.63%
	(2.47)	(1.01)	(-0.11)	(-1.21)	**(-2.57)**	–
Volatility	18.2%	17.75%	17.79%	18.14%	**19.16%**	4.81%
	(0.59)	(-1.65)	(-1.13)	(0.58)	**(3.72)**	–
Sharpe ratio	**0.62**	0.55	0.52	0.48	**0.4**	0.75
	(2.52)	(1.24)	(0.15)	(-1.28)	**(-4.02)**	–
II. Outperformance statistics						
Outperformance	1.89%	0.5%	-0.05%	-0.59%	-1.74%	–
Tracking error	3.19%	2.4%	2.22%	2.29%	2.75%	–
Information ratio	0.59	0.21	-0.02	-0.26	-0.63	–
III. Alpha statistics						
CAPM alpha	**1.94%**	0.73%	0.14%	-0.57%	**-2.24%**	**4.18%**
	(2.42)	(1.53)	(0.26)	(-1.14)	**(-3.56)**	**(3.45)**
Four-factor alpha	**2.2%**	0.64%	0.09%	-0.73%	**-2.21%**	**4.41%**
	(3.05)	(1.44)	(0.18)	(-1.42)	**(-3.74)**	**(3.82)**

Panel B: HY	Q1	Q2	Q3	Q4	Q5	TB
I. Return statistics						
Mean return	4.19%	**4.93%**	3.37%	3.68%	2.26%	1.93%
	(0.67)	**(2.47)**	(-0.79)	(-0.01)	(-1.73)	–
Volatility	**9.24%**	10.13%	10.21%	10.74%	**12.12%**	5.34%
	(-2.06)	(-0.19)	(0.11)	(1.58)	**(2.79)**	–
Sharpe ratio	0.45	0.49	0.33	0.34	**0.19**	0.36
	(1.83)	(1.72)	(-0.75)	(-0.06)	**(-2.54)**	–
II. Outperformance statistics						
Outperformance	0.51%	1.25%	-0.31%	-0.0%	-1.42%	–
Tracking error	2.55%	2.28%	2.19%	2.74%	3.62%	–
Information ratio	0.2	0.55	-0.14	-0.0	-0.39	–
III. Alpha statistics						
CAPM alpha	0.95%	**1.36%**	-0.24%	-0.08%	**-1.96%**	**2.91%**
	(1.78)	**(2.71)**	(-0.62)	(-0.13)	**(-3.02)**	**(3.04)**
Four-factor alpha	1.21%	0.06%	-0.13%	-0.41%	-0.71%	**1.92%**
	(1.81)	(0.14)	(-0.36)	(-0.55)	(-1.26)	**(2.25)**

Panel C: IG	Q1	Q2	Q3	Q4	Q5	TB
I. Return statistics						
Mean return	**1.45%**	1.05%	1.06%	**0.4%**	**0.33%**	1.13%
	(3.78)	(0.98)	(1.31)	**(-2.81)**	**(-2.76)**	–
Volatility	4.33%	**4.17%**	4.45%	4.6%	4.93%	1.38%
	(-1.17)	**(-3.16)**	(0.14)	(1.67)	(1.86)	–
Sharpe ratio	**0.34**	0.25	0.24	**0.09**	**0.07**	0.82
	(4.35)	(1.87)	(1.3)	**(-2.64)**	**(-3.17)**	–
II. Outperformance statistics						
Outperformance	0.59%	0.2%	0.2%	-0.45%	-0.53%	–
Tracking error	0.65%	0.68%	0.69%	0.83%	0.98%	–
Information ratio	0.91	0.29	0.29	-0.54	-0.54	–
III. Alpha statistics						
CAPM alpha	**0.62%**	0.26%	0.21%	**-0.47%**	**-0.61%**	**1.24%**
	(3.93)	(1.65)	(1.36)	**(-2.9)**	**(-2.88)**	**(3.89)**
Four-factor alpha	**0.67%**	0.17%	0.25%	**-0.59%**	**-0.5%**	**1.17%**
	(2.81)	(0.8)	(1.17)	**(-2.26)**	**(-2.03)**	**(2.71)**

This table shows performance statistics of text factors for stocks in the combined MSCI and BMI indexes and corporate bonds in the U.S. Investment Grade and High Yield indexes over the period January 1997 - December 2017. Each month, factor portfolios take equally-weighted long positions in 20% of the stocks and bonds based on text variables such that the highest ranking bonds and stocks are in Q1, the 'next best' 20% are in Q2, and so forth up till the lowest ranking 20% which are in Q5. At time t, the text variables are computed from the text in 10-K and 10-Q filings published in the three preceding months. Mean and volatility are annualized. t-statistics are provided in parentheses for two-sided tests whether the mean, variance, and Sharpe ratio of factor portfolios are different from the corresponding measures of the market portfolio. Bold values indicate statistical significance at the 95% confidence level.

find out what actually drives the performance.

6.3.4 Cumulative return plots

It is common to focus on full sample averages when examining the performance of a strategy. This is useful because it gives insight in the performance of the strategy over a longer period. The rationale is that if a strategy works over the whole sample period, then it should be a good and robust strategy. However, performance of investment strategies is usually not constant, but time varying. Therefore, it is also informative to look at the performance through time, to see whether the strategy works well at all points in time, or whether most gains and losses are realized in specific subperiods. To analyse the performance of the strategy over time, we plot the cumulative log returns and outperformance of the top and bottom portfolio, and the cumulative log returns of the hedge portfolio, over the sample period, in Figure 5.

The plots show that, for all asset classes, the top (bottom) portfolio is above (below) the market at nearly all times. Over the full sample period, the cumulative return of the strategy, for equities, is 199.16% for the top portfolio, compared to 160.50% for the market. For HY, the cumulative return is 120.15% for the top portfolio, compared to 93.94% for the market. For IG, the cumulative return is 32.96% for the top portfolio, compared to 17.23% for the market. The respective hedge portfolios have cumulative returns of 73.34%, 45.49%, and 17.23%. Over the 20 year period, non-changers (changers) have thus consistently outperformed (underperformed) the market for all asset classes. This is, once again, a clear sign that the strategy works. However, the strategy is certainly not flawless. There are some periods of flat and even negative (positive) returns for the top (bottom) portfolios and hedge portfolios. What stands out from the plots is that most of the large upward (downward) movements in the returns of the top (bottom) portfolios and hedge portfolios seem to be concentrated around large downturns in the equity market, denoted by the vertical grey bars, while most of the flat and negative (positive) returns are concentrated in the periods between the downturns. This is a striking finding which suggests that the strategy might work so well because it is able to pick the right companies when most companies experience their largest drawdowns.

6.3.5 Downturn versus non-downturn periods

To test the difference in performance of the composite similarity strategy in non-downturn and downturn periods, we perform separate backtests for both scenarios.

Table 13 shows the four-factor-alpha's of these backtests for the top portfolio, bottom portfolio, and hedge portfolio. The outcome of the test is very clear. The strategy yields positive (negative) and significant alpha for most of the top (bottom) portfolios and hedge

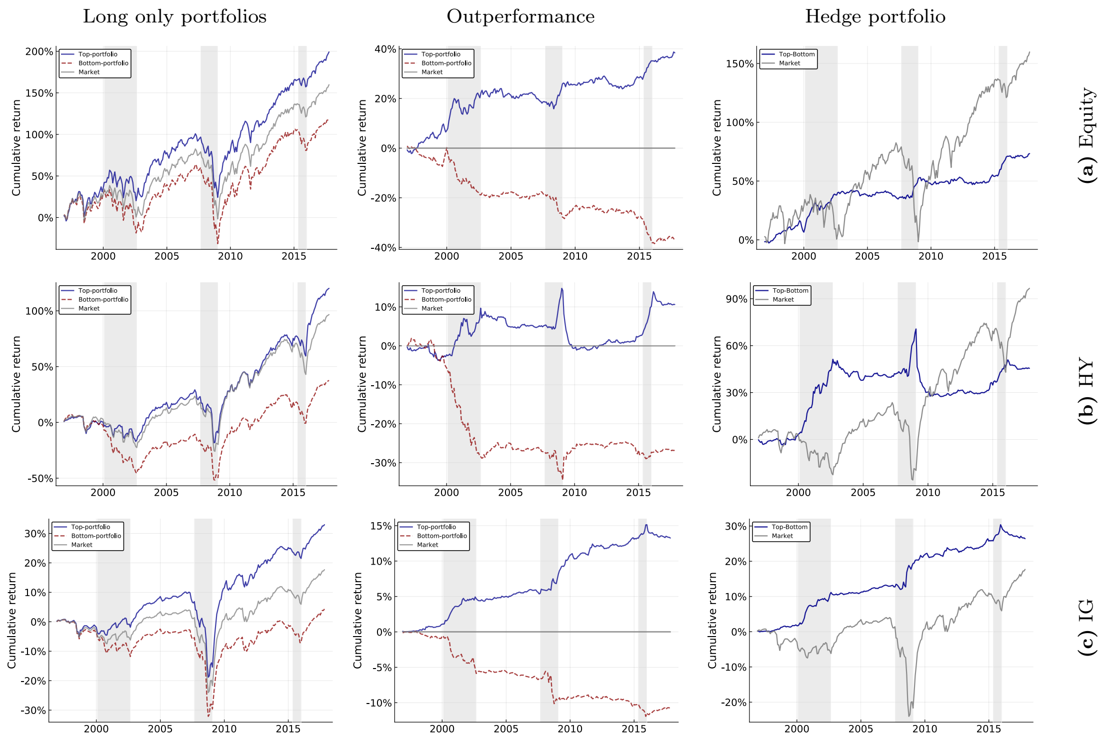


Fig 5: Cumulative returns composite similarity strategy

portfolios in downturns, while there is not a single significant alpha over the non-downturn periods. This confirms that the strategy is able to outperform the market during crashes. Also, the outperformance is quite large, with alpha's of 11.96% (3.45) for equity, 8.94% (3.91) for HY, and 2.75% (2.76) for IG hedging portfolios.

Table 13: Downturn and non-downturn performance

	Non-downturn			Downturn		
	Equity	HY	IG	Equity	HY	IG
Top	0.57%	0.23%	0.15%	**6.16%**	**3.21%**	**1.46%**
	(0.87)	(0.33)	(0.97)	**(2.58)**	**(2.25)**	**(3.35)**
Bottom	-0.26%	-0.61%	-0.14%	**-5.80%**	**-5.74%**	**-1.29%**
	(-0.55)	(-0.74)	(-1.28)	**(-3.59)**	**(-2.95)**	**(-2.17)**
Top-Bottom	0.84%	0.74%	0.29%	**11.96%**	**8.94%**	**2.75%**
	(0.83)	(0.82)	(1.28)	**(3.45)**	**(3.91)**	**(2.76)**

This table contains alpha statistics for the composite similarity strategy in downturn and non-downturn periods. All values are annualized. *t*-statistics are provided in parentheses for two-sided tests whether the alpha's deviate from zero. Bold values indicate statistical significance at the 95% confidence level.

Outperformance during a crash is generally referred to as crisis alpha (Greyserman & Kaminski, 2014). A clear example of crisis alpha is shorting the market when it crashes. Of course, it is very hard to time the market. Therefore, this is generally not advised. The composite similarity strategy on the other hand is seemingly able to generate crisis alpha without the need of any market timing. Crisis alpha is a very sought-after feature for a strategy, since strategies that exhibit it perform well when most other strategies have their worst performance. Because of this, such strategies potentially are excellent diversifiers.

6.3.6 Multi-factor strategy

Individual factors can experience periods of underperformance even though they (are expected to) deliver a premium against the market over time (Ang, 2014). In fact, it is generally believed that these 'bad times' are one of the major reasons why there even are factor premiums in the first place. The periods of underperformance of the most common factors have historically not coincided with each other. So, when one factor is underperforming, other factors may be outperforming. Therefore, in practice, investors are advised not to 'put all their eggs in one basket', but to get exposure to a basket of factors by investing in a multi-factor strategy. Such a strategy ranks stocks and bonds based on a weighted combination of factors instead of a single factor. The factors in a multi-factor portfolio are often given equal weights because factor timing (dynamically allocating to different factors based on a (risk-adjusted) return forecast) has been found to be extremely difficult (Asness, 2016).

To test whether the text in 10-Ks and 10-Qs has additional value for multi-factor portfolios consisting of purely quantitative factors, we perform historical backtests for a multi-factor portfolio consisting of size, value, and momentum and the same multi-factor portfolio with composite similarity included as an additional fourth factor. Table 14 shows the backtest results for both variants. The first variant outperforms the market in all asset classes. The volatility is higher as well, however, this does not offset the increase in return. Hence, the strategy also outperforms in risk-adjusted terms with Sharpe ratios of 0.54, 0.50, and 0.32 for equities, HY, and IG respectively.

Table 14: Multi-factor characteristics of returns Top

	Equity		HY		IG	
	MF	MF+SIM	MF	MF+SIM	MF	MF+SIM
Mean return	10.82%	9.02%	6.23%	5.16%	1.68%	1.54%
Volatility	20.03%	15.03%	12.45%	8.69%	5.23%	3.81%
Sharpe Ratio	0.54	**0.60**	0.50	**0.59**	0.32	**0.40**
	–	**(3.35)**	–	**(3.19)**	–	**(4.41)**

This table contains performance statistics for a multi-factor portfolio of equally-weighted size, value, and momentum factors and the same multi-factor portfolio with composite similarity included as an additional fourth factor. All values are annualized. t-statistics are provided in parentheses for two-sided tests whether the Sharpe ratios between the two strategies deviates from zero. Bold values indicate statistical significance at the 95% confidence level.

The backtest results for the variant with composite similarity included as a fourth factor are also presented in Table 14. The results show that adding composite similarity severely impacts the performance of the multi-factor strategy. On the one hand, the return decreases in all asset classes, while on the other hand, the volatility decreases more. Resulting in Sharpe ratios of 0.60 (3.35), 0.59 (3.19), and 0.40 (4.41) respectively, where the values between parentheses denote t-statistics for a significance test that compares the Sharpe ratios of the two multi-factor strategies. The increase in Sharpe ratio is significant in all asset classes, with t-stats > 3. Adding composite similarity thus robustly improves the risk-adjusted performance of the multi-factor strategy. This shows that including textual factors next to numerical factors in a multi-factor strategy can lead to more robust and better (risk-adjusted) performance.

6.3.7 Robustness checks

We perform a series of robustness checks to ensure that the results for the composite similarity strategy hold more generally. To do so, we perform backtests on the strategy with various different model choices.

Table 17 shows the performance for longer holding-periods (1, 3, 5, 7, and 9 months). Panel A presents the results for equities. The four-factor alpha of the top, bottom, and hedge portfolio decays with the holding period but remains highly significant even after nine months (effectively about one year because of the way the strategies are constructed). This shows that slowly but surely the incremental information of the text in 10-Ks and 10-Qs gets incorporated in the price. Moreover, as the holding period gets longer, new 10-Ks and 10-Qs get published which might make part of the information in old reports redundant. Panel B and C present the results for HY and IG. The alpha of the portfolios again decays with the holding period. For HY (IG), the alpha of the hedge portfolio stays significant until a three (seven) month holding period. So, apparently the information gets incorporated quicker in the corporate bond price than in the stock price of a company.

Table 16 shows the performance for portfolios with weights proportional to a company's market cap for stocks or total amount of debt outstanding for bonds instead of equally-weighted portfolios. The alpha's of the hedge portfolios stay significant for all three asset classes. Hence, the performance of the strategy is not driven by smaller companies. The table also shows the performance for decile portfolios instead of quintile portfolios. The alpha's of the hedge portfolios decrease for equities and HY and even become insignificant. In contrast, for IG, the alpha increases and stays significant. The difference in results for equities compared to the quintile portfolio case indicates that companies in the bottom-of-the-bottom are not necessarily the worst companies. A potential reason for this could be that analysts and investors give most of their attention to these bottom-of-the-bottom companies while neglecting companies with less extreme changes. The results support this hypothesis since for equities the alpha's follow a monotonically decreasing pattern up until the ninth portfolio which has a negative and significant alpha of -3.03% (-3.84). Lastly, the table also shows the performance for portfolios based on only the 10-Ks. To ensure that the sample has enough breath and stays relatively stable over time, we fill the months until the next publication of a 10-K up to a maximum of 13 months for the composite similarity variable. For equities, the alpha's decrease but stay highly significant, while for HY and IG, almost all alpha's become insignificant. The alpha's becoming insignificant for HY and IG is not surprising given the increase in effective holding period of the strategy. The decrease in alpha for all asset classes when only the 10-Ks are used shows that the inclusion of 10-Qs is positive for performance.

6.4 The information content of subsections

Up until now, all analyses have focused on the text in the complete filing. In this section, we will redo the main tests in this thesis for text variables constructed from the text in only the i) MDA, and ii) RF sections of the 10-Ks and 10-Qs. These subsections are amongst the most important parts of the reports and their content has a much more narrow scope than the content in the complete filing. The difference in content can be seen from Figure 6 which shows wordclouds of the text in the complete filing, MDA, and RF for a random selection of reports, where the size of the words indicates the term frequency of the words (i.e. the larger the font the more it occurs in the text).

Table 18 shows the results for the volatility regressions with a single text variable. Many of the variables that were significantly related to future volatility are also significantly related to future volatility when constructed from the text in the MDA and RF. Moreover, the signs of these relations for both subsections are almost all the same as before. Indicating that all previous results for risk prediction have not just been driven by a single subsection of the reports. The only variables that have substantially different effects for the two subsections are the (changes in) sentiment variables. In the MDA, (the change in) polarity is negatively related to future volatility for all three asset classes, while in the RF, the relation is more or less flat. Negative (incremental) information in the RF thus does not affect future volatility. This seems contradicting, since you would expect that more negative information means more risks and therefore higher future volatility. However, managers fear litigation that may result from them failing to disclose a risk that later develops into a negative outcome, even though the outcome might have been highly improbable ex ante. Hence, they tend to adopt a “kitchen sink” approach and disclose a large number of risks, even ones that pose a minimal threat to the company (Chin et al., 2018).

Next, Table 19 shows the results for return regressions with a single text variable. For equities, the similarity variables are significantly related to future return in both the MDA and RF. In contrast, for corporate bonds the similarity variables are not significantly related

(a) Complete filing (b) MDA (c) RF

Fig 6: Wordclouds for text in the a) complete filing, b) MDA, and c) RF of a random selection of reports

to future return. This means that other sections in the report must drive the significant result for the complete filing.

Finally, Table 20 presents four-factor alpha's for the top, bottom, and hedge portfolio of the composite similarity strategy in a historical backtest. For equities, the hedge portfolio earns significant alpha's of 2.91% (2.30) and 5.22% (3.77) in the MDA and RF respectively. For the MDA, this comes mostly from the short leg, while for the RF, it comes from both the long and short leg. This shows that for equities large changes to both the MDA and RF are perceived as bad. Echoing the regression results, the alpha's of the hedge portfolio for corporate bonds are not significant.

7 Conclusion and discussion

This thesis examines the characteristics and variations within the text in 10-K and 10-Q filings over a sample period spanning more than two decades. In line with previous research, we find that (changes in) readability, sentiment, and similarity variables are predictive of future stock price volatility and returns. As a true out-of-sample test, we apply the same methodology to the corporate bond market. Consistent with underreaction due to limited attention of analysts and investors, we find that readability, sentiment, and similarity variables predict future corporate bond price volatility and returns in both the investment grade (IG) and high yield (HY) segments of the U.S. corporate bond market. Altogether, these results show that the informational content of the text in 10-Ks and 10-Qs is not efficiently incorporated in the stock and bond price of a company. We demonstrate that the underreaction to the text in these reports can systematically be exploited by an investment strategy that buys stocks or bonds of companies with few changes compared to previous reports ('non changers') and sells companies with many changes ('changers') as it earns economically and statistically significant (risk-adjusted) returns in both the equity and corporate bond market, even when controlled for quantitative equity and bond factors. The results are not just driven by one particular part of the reports as overall consistent results are found for the Management's Discussion and Analysis (MDA) and Risk Factors (RF) subsections.

The findings in this thesis have implications for the Securities and Exchange Commission (SEC) and for investors in both the stock and coroprate bond market. For the SEC, the post file date predictability of stock and bond prices may be concerning since apparently not all corporate information gets efficiently incorporated in security prices, which is one of their main responsibilities. The SEC acknowledges the need for disclosure reform to battle the information overload that investors face. A recent example of an attempt to improve the effectiveness of the filings is that starting from the beginning of 2018 companies are allowed to add a "fair and accurate" summary to the end of their report.[20] Moreover, on March 20, 2019 the SEC voted to adopt the FAST Act amendment[21] to modernize and simplify disclosure requirements with the intention to "improve the readability and navigability of disclosure documents and discourage repetition and disclosure of immaterial information". For investors, on the other hand, the post file date predictability of stock and bond prices shows that there is value in reading companies' annual and quarterly reports. Besides, by automating the 'reading process' investors can more quickly act on the information in the text of newly filed reports which should give them an edge over investors who (try to) read the reports themselves. In addition, automating the process allows investors to analyse many reports in a short period of time because the process is fast and highly parallelizable, whereas investors who (try to) read the reports themselves already need a considerable amount of time to finish only one report. So, for investors it seems beneficial to start incorporating text analysis for 10-Ks and 10-Qs in their investment process. Maybe not immediately as the only determinant of buys and sells, but for instance as a recommendation system that gives a notification when a report of a company of interest changes by more than $x\%$ or adds (removes) certain keywords compared to previous reports.

[20] https://www.sec.gov/rules/interim/2016/34-77969.pdf

[21] https://www.sec.gov/rules/final/2019/33-10618.pdf

Commemorate that the results in this thesis have been obtained with text variables that are relatively straightforward. In fact, the level of text-understanding that the text variables provide is a fair bit below what humans are capable of (Rudkowsky et al., 2018). Therefore, there is still lots of room for improvement for further research by using more complex models that are better able to capture the nuances and behavioral expressions in text. Examples of candidate models are n-grams, word embeddings, or even a fully fledged state-of-the-art language model like BERT (Devlin et al., 2018). Another promising direction for future research is to focus on other text sources, such as other SEC filings (Form 4, Form 8-K, or Proxy statements) or, for instance, earnings call transcripts, because similar text analysis methods are likely to work there as well.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (Tech. Rep.). National Bureau of Economic Research.

Amel-Zadeh, A., & Faasse, J. (2016). The information content of 10-k narratives: Comparing md&a and footnotes disclosures.

Ang, A. (2014). *Asset management: A systematic approach to factor investing.* Oxford University Press.

Asness, C. S. (2016). The siren song of factor timing.

Bartov, E., & Konchitchki, Y. (2017). Sec filings, regulatory deadlines, and capital market consequences. *Accounting Horizons*, *31*(4), 109–131.

Blitz, D., & Van Vliet, P. (2007). The volatility effect: Lower risk without lower return.

Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, *63*(2-3), 329–357.

Brown, S. V., & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, *49*(2), 309–346.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57–82.

Chin, M. V., Liu, Y., & Moffitt, K. (2018). Voluntary disclosure through the ranking of risk factors in the 10-k.

Chouliaras, A. (2015). The pessimism factor: Sec edgar form 10-k textual analysis and stock returns.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047–1108.

Cohen, L., Malloy, C., & Nguyen, Q. (2018). *Lazy prices* (Tech. Rep.). National Bureau of Economic Research.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*.

Dimson, E., Marsh, P., & Staunton, M. (2017). Factor-based investing: the long-term evidence. *Journal of Portfolio Management*, *43*(5), 15.

Dor, A. B., Dynkin, L., Hyman, J., Houweling, P., van Leeuwen, E., & Penninga, O. (2007). Dtssm (duration times spread). *The Journal of Portfolio Management*, *33*.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data. In *Proceedings of the fifth acm sigkdd international conference on* *knowledge discovery and data mining* (pp. 398–401).

Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis*, *52*(2), 811–836.

Fabozzi, F. J., Pollack, I. M., & Fabozzi, T. D. (1995). *The handbook of fixed income securities*. Irwin Burr Ridge, IL.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34–105.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3–56.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, 387–401.

Givoly, D., Hayn, C., & Katz, S. (2017). The changing relevance of accounting information to debt holders over time. *Review of Accounting Studies*, *22*(1), 64–108.

Grefenstette, G., & Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenisation.

Greyserman, A., & Kaminski, K. (2014). *Trend following with managed futures: The search for crisis alpha*. John Wiley & Sons.

Gunning, R. (1952). The technique of clear writing.

Haesen, D., Houweling, P., & Van Zundert, J. (2013). Residual equity momentum for corporate bonds.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68.

Heij, C., Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Houweling, P., & Van Zundert, J. (2017). Factor investing in the corporate bond market. *Financial Analysts Journal*, *73*(2), 100–115.

Ibbotson, R. G., & Kim, D. Y.-J. (2017). *Liquidity as an investment style: 2017 update* (Tech. Rep.).

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). Asq Press.

Intertrust. (2018). *Changing tides - global private debt market in 2018* (Tech. Rep.).

Israel, R., Palhares, D., & Richardson, S. A. (2016). Common factors in corporate bond and bond fund returns.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, *48*(1), 65–91.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining.* Springer Science & Business Media.

Karapandza, R. (2016). Stock returns and future tense language in 10-k reports. *Journal of Banking & Finance*, *71*, 50–61.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Koller, T., Goedhart, M., Wessels, D., et al. (2010). *Valuation: measuring and managing the value of companies* (Vol. 499). John Wiley and sons.

Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, *15*(5), 850–859.

Ledoit, O., & Wolf, M. (2011). Robust performances hypothesis testing with the variance. *Wilmott*, *2011*(55), 86–89.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, *45*(2-3), 221–247.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, *47*(1), 13–37.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643–1671.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications.* Academic Press.

Newey, W. K., & West, K. D. (1986). *A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.* National Bureau of Economic Research Cambridge, Mass., USA.

Palhares, D., & Richardson, S. A. (2018). (il) liquidity premium in credit markets: A myth?

Patton, A. J., & Timmermann, A. (2008). Portfolio sorts and tests of cross-sectional patterns in expected returns.

Powers, D. M. (1998). Applications and explanations of zipf's law. In *Proceedings of the* *joint conferences on new methods in language processing and computational natural language learning* (pp. 151–160).

Ravazzolo, F. (2007). *Forecasting financial time series, using model averaging* (No. 415). Rozenberg Publishers.

Rogers, J. L., Van Buskirk, A., & Zechman, S. L. (2011). Disclosure tone and shareholder litigation. *The Accounting Review*, *86*(6), 2155–2183.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140–157.

SEC. (2013). *Report on review of disclosure requirements in regulation s-k.* Retrieved from `https://www.sec.gov/news/studies/2013/reg-sk-disclosure-requirements-review.pdf`

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Singhal, A., et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35–43.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139–1168.

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, *39*(5), 4729–4739.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

A Tables and Figures

Table 15: Section definitions

Item	Name
Part I	
1	Business
1A	Risk Factors
1B	Unresolved Staff Comments
2	Properties
3	Legal Proceedings
4	Mine Safety Disclosures
Part II	
5	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
6	Selected Financial Data
7	Management's Discussion and Analysis of Financial Condition and Results of Operations
7A	Quantitative and Qualitative Disclosures About Market Risk
8	Financial Statements and Supplementary Data
9	Changes in and Disagreements With Accountants on Accounting and Financial Disclosure
9A	Controls and Procedures
9B	Other Information
Part III	
10	Directors, Executive Officers and Corporate Governance
11	Executive Compensation
12	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
13	Certain Relationships and Related Transactions, and Director Independence
14	Principal Accounting Fees and Services
Part IV	
15	Exhibits, Financial Statement Schedules
16	Form 10–K Summary

(a) 10-K

Item	Name
Part I	
1	Financial Statements
2	Management's Discussion and Analysis of Financial Condition and Results of Operations
3	Quantitative and Qualitative Disclosures About Market Risk
4	Controls and Procedures
Part II	
1	Legal Proceedings
1A	Risk Factors
2	Unregistered Sales of Equity Securities and Use of Proceeds
3	Defaults Upon Senior Securities
4	Mine Safety Disclosures
5	Other Information
6	Exhibits

(b) 10-Q

Table 16: Robustness tests

	Value-weighted			Decile portfolios			Only 10-Ks		
	Equity	HY	IG	Equity	HY	IG	Equity	HY	IG
Top	**3.10%**	**1.42%**	**0.75%**	**2.34%**	0.53%	**0.78%**	**1.64%**	-0.76%	0.03%
	(3.37)	**(2.41)**	**(3.28)**	**(2.37)**	(0.88)	**(2.42)**	**(2.58)**	(-1.19)	(0.15)
Bottom	-1.14%	-0.77%	**-0.76%**	-0.56%	**-1.47%**	**-1.16%**	**-1.18%**	0.66%	**-0.48%**
	(-1.43)	(-0.87)	**(-2.00)**	(-0.50)	**(-2.55)**	**(-2.16)**	**(-2.07)**	(0.92)	**(-2.49)**
Top-Bottom	**4.24%**	**2.18%**	**1.52%**	2.90%	2.00%	**1.94%**	**2.82%**	-1.42%	0.51%
	(3.18)	**(2.37)**	**(3.04)**	(1.76)	(1.94)	**(2.40)**	**(2.88)**	(-1.20)	(1.54)

This table contains alpha statistics for the composite similarity strategy with i) value-weighted portfolios, ii) decile portfolios, and iii) portfolios based on only the 10-Ks. All values are annualized. t-statistics are provided in parentheses for two-sided tests whether the alpha's deviate from zero. Bold values indicate statistical significance at the 95% confidence level.

Table 17: Holding period analysis

Panel A: Equity					
$H =$	1	3	5	7	9
Top	**1.82**	**1.75**	**1.78**	**1.45**	**1.40**
	(2.79)	**(2.63)**	**(2.9)**	**(2.53)**	**(2.54)**
Bottom	**-1.32**	**-1.22**	**-1.13**	**-1.01**	**-0.90**
	(-2.37)	**(-2.55)**	**(-2.42)**	**(-2.33)**	**(-2.13)**
Top-Bottom	**3.14**	**2.97**	**2.91**	**2.46**	**2.30**
	(3.16)	**(2.96)**	**(2.98)**	**(2.72)**	**(2.63)**
Panel B: HY					
$H =$	1	3	5	7	9
Top	1.21	0.78	0.42	0.27	-0.00
	(1.81)	(1.44)	(0.89)	(0.66)	(-0.01)
Bottom	-0.71	-0.96	-0.93	**-1.09**	**-1.20**
	(-1.26)	(-1.70)	(-1.61)	**(-2.10)**	**(-2.50)**
Top-Bottom	**1.92**	**1.64**	1.36	1.36	1.19
	(2.25)	**(2.03)**	(1.55)	(1.79)	(1.73)
Panel C: IG					
$H =$	1	3	5	7	9
Top	**0.51**	0.33	0.31	0.29	0.23
	(2.62)	(1.79)	(1.68)	(1.58)	(1.28)
Bottom	-0.49	**-0.45**	-0.38	-0.35	-0.25
	(-1.86)	**(-2.31)**	(-1.86)	(-1.75)	(-1.37)
Top-Bottom	**1.00**	**0.78**	**0.69**	**0.64**	0.48
	(2.52)	**(2.46)**	**(2.14)**	**(2.01)**	(1.63)

This table contains alpha statistics for the composite similarity strategy with holding periods of 1, 3, 5, 7, and 9 months. All values are annualized. t-statistics are provided in parentheses for two-sided tests whether the alpha's deviate from zero. Bold values indicate statistical significance at the 95% confidence level.

Table 18: Regressions results for future 1-year volatility in MDA and RF

	Equity		Credits HY		Credits IG	
	MDA	RF	MDA	RF	MDA	RF
LEN	**0.81**	**1.33**	1.14	2.25	**1.65**	0.33
	(2.47)	**(3.28)**	(0.95)	(1.13)	**(2.12)**	(0.30)
ΔLEN	**1.06**	**1.20**	**2.21**	2.50	**1.57**	**2.06**
	(5.11)	**(3.45)**	**(2.54)**	(1.64)	**(2.51)**	**(2.26)**
POL	**-0.90**	**0.87**	**-2.88**	-1.48	**-1.96**	-0.67
	(-3.69)	**(2.50)**	**(-2.94)**	(-1.07)	**(-2.75)**	(-0.67)
POS	0.17	**1.09**	-1.31	**-3.29**	**-2.89**	**-2.03**
	(0.74)	**(2.74)**	(-1.16)	**(-2.03)**	**(-3.94)**	**(-2.28)**
NEG	**1.25**	0.37	1.58	-2.29	0.32	-0.21
	(3.97)	(1.07)	(1.41)	(-1.43)	(0.42)	(-0.21)
ΔPOL	**-1.11**	-0.20	**-1.98**	0.76	**-2.45**	0.13
	(-5.44)	(-0.67)	**(-2.55)**	(0.50)	**(-4.07)**	(0.17)
ΔPOS	-0.25	0.14	-0.09	0.05	**-1.13**	0.85
	(-1.35)	(0.41)	(-0.10)	(0.03)	**(-2.09)**	(1.00)
ΔNEG	**1.13**	0.14	**1.82**	-0.46	**2.19**	0.06
	(5.57)	(0.48)	**(2.10)**	(-0.29)	**(3.65)**	(0.07)
COS	**-0.59**	-0.26	-1.64	**-3.01**	**-1.70**	**-2.27**
	(-2.64)	(-0.78)	(-1.78)	**(-2.2)**	**(-2.48)**	**(-2.41)**
JAC	-0.11	-0.50	**-2.08**	**-3.13**	**-1.83**	**-2.48**
	(-0.48)	(-1.43)	**(-2.09)**	**(-2.33)**	**(-2.64)**	**(-2.67)**

This table contains regression results for panel regressions of future 3-month return on text variables and control variables. The text variables are constructed from the text in the i) MDA, and ii) RF. In each regression only a single text variable is included. Size, value, and momentum are included as control variables. Moreover, beta and DTS-beta are respectively included as additional regressors for the equity and corporate bond regressions. Robust $t-$statistics are provided in parentheses for a test on the difference from zero for the regression coefficients. Bold values indicate statistical significance at the 95% confidence level.

Table 19: Regressions results for future 3-month return in MDA and RF

	Equity		Credits HY		Credits IG	
	MDA	RF	MDA	RF	MDA	RF
LEN	-0.19	0.19	0.48	0.24	0.06	0.2
	(-0.61)	(0.47)	(1.38)	(0.43)	(0.46)	(0.69)
ΔLEN	-0.22	-0.28	-0.15	0.13	0.12	0.1
	(-0.89)	(-0.81)	(-0.45)	(0.24)	(1.13)	(0.44)
POL	0.35	-0.17	-0.04	-0.22	-0.09	-0.16
	(1.75)	(-0.49)	(-0.15)	(-0.56)	(-0.86)	(-1.23)
POS	0.43	0.41	0.18	-0.14	0.06	-0.06
	(1.9)	(1.01)	(0.8)	(-0.27)	(1)	(-0.4)
NEG	-0.43	0.56	0.25	0.09	0.11	0.12
	(-1.2)	(1.77)	(0.67)	(0.19)	(0.86)	(0.97)
ΔPOL	0.02	-0.43	-0.13	-0.29	0.02	-0.12
	(0.07)	(-1.46)	(-0.43)	(-0.66)	(0.32)	(-0.73)
ΔPOS	-0.21	**-0.74**	0	0.45	0.05	-0.13
	(-1.01)	**(-2.52)**	(-0.01)	(0.82)	(0.88)	(-0.99)
ΔNEG	-0.11	0.19	0.03	0.56	0.01	0.09
	(-0.47)	(0.71)	(0.1)	(1.17)	(0.16)	(0.51)
COS	**0.49**	**0.86**	0.01	-0.71	0.08	-0.03
	(2.36)	**(2.95)**	(0.04)	(-1.39)	(1.25)	(-0.18)
JAC	0.46	**0.82**	0.18	-0.68	0.04	-0.06
	(1.92)	**(2.88)**	(0.57)	(-1.49)	(0.63)	(-0.29)

This table contains regression results for panel regressions of future 1-year volatility on text variables and control variables. The text variables are constructed from the text in i) the MDA, and ii) the RF. In each regression only a single text variable is included. Size, value, and momentum are included as control variables. Moreover, past 1-year volatility and DTS are respectively included as additional regressors for the equity and corporate bond regressions. Robust $t-$statistics are provided in parentheses for a test on the difference from zero for the regression coefficients. Bold values indicate statistical significance at the 95% confidence level.

Table 20: Performance composite similarity strategy in MDA and RF

	MDA			RF		
	Equity	HY	IG	Equity	HY	IG
Top	0.93	**1.33**	-0.08	**1.74**	-0.82	0.3
	(1.09)	**(2.65)**	(-0.34)	**(2.31)**	(-1.07)	(1.17)
Bottom	**-1.97**	0.18	-0.07	**-3.48**	1.04	-0.55
	(-3.43)	(0.27)	(-0.35)	**(-3.55)**	(0.78)	(-1.04)
Top-Bottom	**2.91**	1.15	-0.01	**5.22**	-1.86	0.84
	(2.3)	(1.63)	(-0.01)	**(3.77)**	(-1.11)	(1.26)

This table contains alpha statistics for the composite similarity strategy based on the text in the i) Management Discussion and Analyses (MDA) section, and ii) Risk Factors (RF) section. All values are annualized. t-statistics are provided in parentheses for two-sided tests whether the alpha's deviate from zero. Bold values indicate statistical significance at the 95% confidence level.

B Robust performance hypothesis testing

Let there be two investment strategies i and j with excess (log) returns $r_{t,i}$ and $r_{t,j}$ for $t = 1, \ldots, T$. Assume that the observations $(r_{1,i}, r_{1,j})', \ldots, (r_{T,i}, r_{T,j})'$ are stationary.[22] Then, the returns of the strategies follow a bivariate distribution that is constant over time with mean vector $\boldsymbol{\mu} = (\mu_i, \mu_j)'$ and covariance matrix $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{smallmatrix}\right)$. The moments are unbiasedly estimated by their sample analogs:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{r}_t, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{r}_t - \widehat{\boldsymbol{\mu}})(\boldsymbol{r}_t - \widehat{\boldsymbol{\mu}})'. \tag{24}$$

B.1 Test on difference of means

A test on the difference in means of the two strategies comes down to testing the following hypotheses:

$$H_0 : \mu_i - \mu_j = 0 \quad \text{vs.} \quad H_1 : \mu_i - \mu_j \neq 0. \tag{25}$$

This test can be performed by the following regression

$$r_{i,t} - r_{j,t} = \Delta + \varepsilon_t, \tag{26}$$

where Δ is a constant and ε_t is an idiosyncratic error term. The resulting estimate $\widehat{\Delta}$ is the estimated sample difference between the means and $s_{\widehat{\Delta}}$ is the corresponding Newey-West corrected standard error (Newey & West, 1986). A test on the difference in means can than be formulated as

$$t = \frac{\widehat{\Delta}}{s_{\widehat{\Delta}}/\sqrt{T}}, \tag{27}$$

with corresponding p-value

$$\widehat{p} = 2t_{cdf}\left(-\frac{|\widehat{\Delta}|}{s(\widehat{\Delta})}\right). \tag{28}$$

B.2 Test on difference of variances

Define the difference in variances of the two strategies as

$$\Delta = \log(\Theta) = \log(\sigma_i^2) - \log(\sigma_j^2).^{23} \tag{29}$$

Testing the difference in variances of the strategies then comes down to testing the following hypotheses:

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0. \tag{30}$$

[22] A bivariate process $\boldsymbol{X}_t = (X_{t1}, X_{t2})'$ for $t = 1, \ldots, T$ is said to be (weakly) stationary if its first and second moments are time invariant: 1) $E(\boldsymbol{X}_t) = E(\boldsymbol{X}_{t-1}) = \boldsymbol{\mu}$ for all t, 2) $Var(X_{ti}) < \infty$ for all t and $i = 1, 2$, and 3) $E((\boldsymbol{X_t} - \boldsymbol{\mu})(\boldsymbol{X_{t+k}} - \boldsymbol{\mu})') = \boldsymbol{\Gamma}_k$ for all pairs (t, k).

[23] Alternatively, Δ could be defined as $\Delta = \sigma_i^2 - \sigma_j^2$, however, that would lead to an inference method with inferior finite-sample properties (Ledoit & Wolf, 2011)

Δ can be estimated by plugging in the usual sample moments $\widehat{\sigma}_i^2$ and $\widehat{\sigma}_j^2$ so that

$$\widehat{\Delta} = \log(\widehat{\sigma}_i^2) - \log(\widehat{\sigma}_j^2) \tag{31}$$

Now, let $\gamma_i = E(r_{1,i}^2)$ and $\gamma_j = E(r_{1,j}^2)$ be the uncentered second moments and $\widehat{\gamma}_i$ and $\widehat{\gamma}_j$ their respective sample counterparts. Further, define $v = (\mu_i, \mu_j, \gamma_i, \gamma_j)'$ and $\widehat{v} = (\widehat{\mu}_i, \widehat{\mu}_j, \widehat{\gamma}_i, \widehat{\gamma}_j)'$. Then, the difference in variances and its sample estimate can be rewritten as $\Delta = f(v)$ and $\widehat{\Delta} = f(\widehat{v})$, where

$$f(a, b, c, d) = \log(c - a^2) - \log(d - b^2). \tag{32}$$

Now, assume that $\sqrt{T}(\widehat{v} - v) \xrightarrow{d} N(0, \Psi)$ where $\xrightarrow{d}$ denotes convergence in distribution and Ψ is an unknown symmetric positive-definite matrix.[24] The Delta method then implies that

$$\sqrt{T}(\widehat{\Delta} - \Delta) \xrightarrow{d} N(0, \nabla' f(v) \Psi \nabla f(v)), \tag{33}$$

where

$$\nabla' f(a, b, c, d) = \left(-\frac{2a}{c - a^2}, \frac{2b}{(d - b^2)}, \frac{1}{(c - a^2)}, -\frac{1}{(d - b^2)} \right). \tag{34}$$

Ledoit and Wolf (2008) propose a heteroskedasticity and autocorrelation robust (HAC) kernel estimator for the limiting covariance matrix Ψ. See their paper for a detailed description of the estimation steps. Given the estimator $\widehat{\Psi}$, the standard error $s(\widehat{\Delta})$ is obtained as

$$s(\widehat{\Delta}) = \sqrt{\frac{\nabla' f(\widehat{v}) \widehat{\Psi} \nabla f(\widehat{v})}{T}}. \tag{35}$$

Then, combined with the asymptotic normality, a two sided p-value for the null hypothesis $H_0 : \Delta = 0$ is given by

$$\widehat{p} = 2\Phi \left(-\frac{|\widehat{\Delta}|}{s(\widehat{\Delta})} \right), \tag{36}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

B.3 Test on difference of Sharpe ratios

Define the difference in Sharpe ratios of the two strategies as

$$\Delta = SR_i - SR_j = \frac{\mu_i}{\sigma_i} - \frac{\mu_j}{\sigma_j}. \tag{37}$$

Testing the difference in Sharpe ratios then comes down to testing the following hypotheses:

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0. \tag{38}$$

[24] This relation holds under mild regularity conditions (e.g. existence of the fourth moment under i.i.d.)

Δ can be estimated by plugging in the usual sample moments $\widehat{\mu}_i$, $\widehat{\mu}_j$ and $\widehat{\sigma}_i^2$, $\widehat{\sigma}_j^2$ so that

$$\widehat{\Delta} = \widehat{SR}_i - \widehat{SR}_j = \frac{\widehat{\mu}_i}{\widehat{\sigma}_i} - \frac{\widehat{\mu}_j}{\widehat{\sigma}_j}, \tag{39}$$

Now, analogously to before in Section B.2, let $\gamma_i = E(r_{1,i}^2)$ and $\gamma_j = E(r_{1,j}^2)$ be the uncentered second moments and $\widehat{\gamma}_i$ and $\widehat{\gamma}_j$ their sample counterparts. Further, define $v = (\mu_i, \mu_j, \gamma_i, \gamma_j)'$ and $\widehat{v} = (\widehat{\mu}_i, \widehat{\mu}_j, \widehat{\gamma}_i, \widehat{\gamma}_j)'$. Then, the difference in Sharpe ratios and its sample estimate can be rewritten as $\Delta = f(v)$ and $\widehat{\Delta} = f(\widehat{v})$, where

$$f(a, b, c, d) = \frac{a}{\sqrt{c - a^2}} - \frac{b}{\sqrt{d - b^2}}. \tag{40}$$

Again, assume that $\sqrt{T}(\widehat{v} - v) \xrightarrow{d} N(0, \Psi)$. The Delta method then implies that

$$\sqrt{T}(\widehat{\Delta} - \Delta) \xrightarrow{d} N(0, \nabla' f(v) \Psi \nabla f(v)), \tag{41}$$

where

$$\nabla' f(a, b, c, d) = \left(\frac{c}{(c - a^2)^{1.5}}, -\frac{b}{(d - b^2)^{1.5}}, -\frac{1}{2}\frac{a}{(c - a^2)^{1.5}}, \frac{1}{2}\frac{b}{(d - b^2)^{1.5}} \right). \tag{42}$$

HAC inference is now performed similar as before.