Abstract

In recent times, large scale textual analysis was started to be deployed in finance and accounting analysis. Among others, an established use case happens to be the application of natural language processing for volatility forecasting. In this thesis, I extend existing dictionary-based methodologies in the field of textual sentiment analysis by introducing a term-weighting scheme which is based on the impact of words on past volatility. Using a sample of 46,384 corporate 10-K filings, the learned term weights are aggregated to "sentiment scores" for out-of-sample 10-K reports. It is found that – while negative and positive tone embedded in the 10-K do help in explaining post-filing volatility of equity returns – other textual aspects such as assertiveness/uncertainty in the management's writing style, focus on financial topics, or document readability appear to have insignificant importance for volatility forecasting purposes or carry influential power only in certain years of the test period (2013 - 2017). This notion is confirmed in robustness checks which incorporate variance forecasts from powerful conventional time-series models: after including potent volatility forecasts from the GARCH model family, with the exception of positive and negative sentiment the textual contents embedded in 10-K filings fall short in providing value added for prediction of realized stock return volatility.

Contents

1 Introduction

Consider the following situation: investor I is offered a choice between two alternatives. In the first scenario, a dice will be rolled and - depending on the number of eyes displayed - a certain amount will be paid to the investor. In case only one pip appears, she receives 100 USD, if two pips appear she will get 200 USD, and so on. Assuming the dice is fair, her expected pay-off from this game will be 350 USD. If the second alternative would be a guaranteed payment of 350 USD, would the investor be willing to take the gamble and roll the dice?

Most (behavioural) finance research would answer this question with "no", with the main argument being similar to: the *rational* investor does not only care about the average pay-off she will receive (which would be the same in both alternative choices of the game), she is also interested in the riskiness as to which this pay-off will occur. Assuming that the average investor will perceive the expected deviation from the mean payment as "riskiness" (or as it shall be called henceforth: "volatility"), the second option is clearly preferred; the monetary outcome is secured and is not connected to any uncertainty about the payment of the game whatsoever[1]. In other words, as Markowitz (1952, p. 77) pointed out in his pioneering paper about portfolio theory: "the investor does (or should) consider expected return a desirable thing and variance of return an undesirable thing."

Being a variable clearly of interest to investors, the task of forecasting the volatility of financial time series experienced a true "boom" in the past few decades, in both theoretical as well as empirical research, especially after the publication of two seminal papers by Engle (1982) and Bollerslev (1986), who introduced the (G)ARCH models and were the first ones to model conditional variance series as time series. And also practitioners acknowledge the importance of return variance in real-world financial applications, the most famous probably being derivatives pricing, portfolio selection, and risk management. The latter also increased awareness of regulators with regards to monitoring asset volatility.

Moreover, volatility has become to be considered an "asset" per se, in the sense that one can speculate on how it will develop in the future. The most famous means to invest in volatility as an asset is using index-tracking instruments which mimick the Chicago Board Options Exchange SPX Volatility Index (usually referred to as VIX). The VIX provides estimates about general market volatility by combining option-implied volatilities at different strikes for single stocks and aggregating them in a weighted-average method. A very recent real-world example demonstrates the importance in an accurate prediction of volatility for investment purposes: in early February 2018, the VIX increased by more than an unprecedented 100 per cent in a single day. Some exchange-traded products, such

[1] In this game, the first alternative (gamble) exhibits a standard deviation of around 171 USD, while the second alternative (safe payment) has a standard deviation of zero.

as the "Inverse VIX Short-Term exchange-traded note", issued by Swiss bank Credit Suisse and accordingly named XIV, who bet on the "calmness" of markets, lost a substantial amount of value due to the spike in market volatility. The XIV, for instance, closed almost 80 percent down, leaving the issuer of the securities to experience a siginficant 8.5 percent drop in its share price the subsequent day (Franck 2018).

Such examples raise the question on what the drivers of an financial asset's volatility are. Focussing on public equity markets, many factors have been used in many models in the attempt to explain volatility. The most "powerful" proved to be past realizations of volatility; a fact that well describes the empirically observed auto-regressive nature of return variance. It is indeed a "stylized fact" that volatility tends to occur in so-called *clusters*. This circumstance also strongly connects volatility modelling with time series analysis and forecasting, with the main workhorse being the usage of lagged observations of variance in order to explain it "today" (and in the future, as well). Besides past return series, other authors (such as Paye (2012)) have attempted to add further financial and macroeconomic factors in their predictions (e.g., GDP fluctuations or inflation as explanatory indicators); however, there is little consensus in the literature about the usefulness of such variables in volatility modelling (Mittnik et al. 2015, p. 2). Another driver of volatility has shown to be the so-called (option-price-) implied volatility for single stocks, or, on an aggregate level, the VIX for market-level volatility (cf. Mittnik et al. (2015) and referenced literature therein).

More recently, some authors have also attempted to include what can be called "soft" or "qualitative" information in their (volatility) models; very often such variables are used in addition to quantitative data (such as historical time series in prices, returns, or volumes, financial statement and balance sheet metrics, etc.). The latter are "hard" inputs in the sense that they are observable and verifiable, while qualitative information by nature is more subjective and imprecise (Loughran and McDonald 2016). The largest part of such soft information content is available in the form of text. However, as Das (2014, p. 146) points out, "until recently, financial analysis was just based on numbers. Usage of text required human coding of attributes into numerical form before yielding to analysis". Having available computer machines with larger storage capacity, faster processing power and appropriate software to analyze textual inputs, the applications of text mining and processing in the domain of finance and accounting have increased tremendously since the 1990s (Kumar and Ravi 2016). Representatively of this rising attention towards textual information, well-known economist Hal Varian, when reviewing the paper of Antweiler and Frank (2004) in his New York Times column in 2004, stated:

> In the 1970's we saw the rise of Wall Street quantitative analysts. Then came program trading. Perhaps computational linguistics and textual data mining will become the new hot technologies in financial economics. (Varian (2004))

While the terminology of "text mining" is often used interchangeably with phrases such as computational linguistics, content analysis, natural or statistical language processing, text analytics, information retrireval, stylometrics, etc. (Loughran and McDonald 2016), the basic concept for the purposes of this thesis can be defined as follows:

> Text mining is the large-scale, automated processing of plain text language in digital form to extract data that is converted into useful quantitative or qualitative information. (Das (2014))

Referring to a meta-study by Kumar and Ravi (2016), the applications of text mining techniques within the field of research in the business and economics domain are broad; specifically, they categorize them very broadly into four buckets: "FOREX rate prediction, stock market prediction, customer relationship management (CRM) and cyber security" (Kumar and Ravi 2016, p. 128). However, text mining in the business domain is no longer a purely academic undergoing; real-world applications, for instance, can include the measurement of customer satisfaction and churn likelihood by mining unstructured reviews of clients, the development of early-warning systems for banks that incorporate sentiment information in estimation of default probabilities (Kremer et al. 2013), or the establishment of an internal communication and monitoring tool for companies so to measure employee mood, motivation, or satisfaction (Heires 2015). Moreover, the range of applications is very likely to increase in the future, as "it's not hard to see how almost any business could eventually reap rewards from the ability to comb through the writings of millions of people to identify coming desires and/or needs in entertainment, food, travel, retail – pretty much anything, in both the consumer and B2B space" (Belsky 2012).

Furthermore, irrespective of the actual application framework, the universe of textual sources which can be tapped is very large; a non-exhaustive list might range from brokerage/analyst reports, minutes of board or committee meetings, **10-K reports** or other filings and prospectuses with the SEC, corporate announcements, customer reviews and forum posts, news articles, social media and blog posts, e-mails, conference call logs, and many more. 10-K* filings of US-companies with the United States Securities and Exchange Commission (henceforth abbreviated using the common acronym SEC), highlighted in bold face, will be the textual corpus under consideration in this thesis[2], with the final sample containing a rich collection of 46,483 filings from the time period

[2] I will henceforth use the *-notation to denote all variants of 10-K filings. The filing types occurring in the sample of this thesis encompass the following: 10-K, 10-K-A, 10-K405, 10-K405-A, 10-KSB, 10-KSB-A, 10-KT, 10-KT-A, 10KSB, 10KSB-A, 10KSB40, 10KSB40-A, 10KT405 and 10KT405-A. The filing variants 10-K405, 10-K405-A, 10KSB40, 10KSB40-A, 10KT405 and 10KT405-A reports were only allowed up till 2003 and indicate a check-the-box rule (regarding the so-called Regulation S-K or S-B Item 405). KSB reports were eligible for small businesses up till 2009. The interjectional letter `T` indicates transition reports, while addendum `-A` indicates amendments to previously filed reports. For details regarding 10-K types, refer to `https://help.edgar-online.com/edgar/formtypes.asp` (accessed on 07/30/2018).

1999-2017. However, having at hand such a large amount of textual input might bring up problems and obstacles on its own. In this context, Kremer et al. (2013) point out from a practitioner's point of view:

> Enormous quantities of textual information are available [...] and it is notoriously difficult to use. [...] The challenges of mining this information and separating the signal from the noise are substantial. [...] A database with news articles on about 1,000 companies easily exceeds 20 GB, orders of magnitude more than a financial database on these companies. Storing this much data is not difficult, but any kind of statistical analysis becomes an "overnight job", even with optimized algorithms and systems. [...] Second, textual data are unstructured. While it is relatively easy to analyze financial data in a statistical way — figures become meaningful at a certain size and in relation to sample averages — texts are a priori meaningless to a computer. There are no standard or statistical procedures for a machine to analyze and interpret texts. [...] Third, texts are often ambiguous. [...] In fact, almost all the semantic difficulties of written language pose immense problems for machines.

When it comes to the techniques available for analyzing textual contents for financial purposes, a lot of methodologies are available, while the most common stream of activity in this domain can be subsumed under the umbrella of "sentiment analysis". Heires (2015) defines it as follows:

> Sentiment analysis studies the mood, opinions and attitudes expressed in written text. It aims to discover the emotions behind words in order to determine whether a communication suggests a positive, negative or neutral sentiment.

It is the goal of this thesis to combine volatility forecasting with large-scale textual analysis; more specifically, the textual sentiment embedded in 10-K* filings will be extracted in an innovative fashion (sentiment term weighting on the basis of past volatility impact) and then used as an additional explanatory factor in the attempt to predict realized variance after the filing of annual reports. This work thereby contributes to existing literature in two dimensions: research connected to the field of volatility prediction and forecasting as well as textual analysis in finance and accounting. With respect to the first stream of literature, a cross-sectional analysis of 46,483 corporate 10-K filings brings to light that positive and negative sentiment embedded in the reports add incremental value in predicting post-filing realized volatility above a simple past predictor (pre-filing realized volatility) as well as conventional time-series models from the GARCH family. Conversely, more granulated linguistic aspects such as assertiveness, uncertainty, litigiousness, readability, or usage of

financial keywords, provide little value added in explaining post-filing realized volatility. The second contribution of this thesis concerns language analysis in the empirical finance domain. In this aspect, a new term-weighting method was introduced: instead of using raw term frequencies or established corpus-related weighting schemes, I applied a market-based methodology, smoothing term counts based on their contribution in explaining volatility after past 10-K filings. Most findings revealed in this work are robust across multiple dimensions, such as choice of training/test sample split, alternative volatility proxies, or selection of term-weighting scheme. One special side-note that was discovered in robustness checks further sets this work in contrast to research results related to equity returns: as was revealed in an extensive analysis, volatility models from the GARCH-family seem to do an accurate job in forecasting, while such precision is not present in models for return series. Thus, textual analysis on return series is leaving "more room for incremental improvement" coming from qualitative and potentially noisy input features such as text tonality and sentiment.

The remainder of this thesis is structured as follows: section 2 provides an overview about existing literature and theoretical foundations of volatility forecasting and textual analysis in the finance and accounting domain. Section 3 provides an overview of how 10-K* filings can be mined in a vector-space framework and in this context also presents the first block of my research framework, the methodology of volatility-impact-based term weighting. Section 4 introduces the second pillar of the research design, i.e., presents the hypotheses which connect the document sentiment scores (and other control variables) with post-filing volatility and describes the econometric model used. Section 5 gives insights about the data collection and cleaning procedures as well as statistical descriptions about the final sample. Section 6 provides the uni- and multivariate estimations and a discussion of the results. Section 7 offers some thoughts and examinations regarding result robustness. Section 8 draws final conclusions and offers impulses for future research, especially with regards to alternative research design specifications.

2 Literature Review and Theoretical Background

In this section I will provide an overview about influential literature in the field of textual analysis in finance and accounting, with a particular focus on contributions focussing on sentiment analysis. Moreover, I will elaborate on publications related to the readability of financial documents. Finally, a short summary about literature that uses corporate filings for text analysis purposes is provided.

2.1 Text Mining and Natural Language Processing in Finance and Accounting

As Yukselturk and Tucker (2015) evidence, research based on content analysis in the areas of accounting and finance has gained significant importance recently, with growing literature available in the fields of corporate disclosure and financial reporting as well as analyst reports; in the same context they also report a "growing body of research on the impact of tone or sentiment on asset prices and returns" (Yukselturk and Tucker 2015, p. 871). Excellent meta-studies on the use of textual analysis within finance and accounting are given in Kearney and Liu (2014)[3], Guo et al. (2016) as well as Loughran and McDonald (2016). As the latter point out, the idea of mining textual contents to search for patterns and information has a long tradition and reaches far back into history, ranging from biblical keyword analysis to the dissection of the rhetorical content in political speeches. In more recent times, both large increases in computing power as well as availability of textual material have also encouraged researchers from the finance and accounting disciplines to apply textual analysis in these domains. Loughran and McDonald (2016, p. 1188), with regard to this trend, trenchantly point out that "the online availability of news articles, earnings conference calls, Securities and Exchange Commission (SEC) filings, and text from social media provide ample fodder for applying the technology."

A generic research design of finance and accounting studies which encompass textual analysis is schematically depicted in Figure 1: starting with the textual "underlying", one usually seeks to extract some output from a collection of documents (often referred to as *corpus*). The desired outputs are various (such as sentiment scores, readability measures, similarity metrics, topic categories, document rankings, and many more); likewise, the methods to extract such variables from text are manifold and can include counting the appearance of certain keywords or chain of words (so called *n-grams*), semantic decompositions, the classification of sentences, paragraphs, or documents, or more sophisticated machine learning algorithms such as support vector methods or neural

[3] Especially their Table 1 and Table 3 provide a good overview of influential contributions, grouped by the source of textual inputs used, and the main findings, respectively (Kearney and Liu 2014, pp. 3, 12).

networks. These text-related outputs are then usually connected with quantitative information, whereas tools range from established statistical and econometric procedures (mostly linear or logistic regression) to more advanced fusion methods that involve algorithms from machine learning.

In general, however, natural language processing in the financial domain brings up several obstacles and difficulties for the researcher. A fundamental question is related to an essential trade-off between signal and noise when choosing the research design; in other words, although analyses sometimes use simplifying assumptions[4], the attempt to better capture meaning and context also from a syntactic and semantic point of view can in some cases do more harm than good, because one simply adds more noise to the model that shall not be mistaken for signal (Loughran and McDonald 2016).

Further issues, which affect a lot of studies, are data availability, retrieval issues such as downloadability, document structure, and the subsequent parsing process. For some corpora, no collections or databases are available and require download by hand. Moreover, the downloaded files are often not available in the required format or are not machine-readable[5], or require at least a minimum amount of parsing. This imposes large practical problems, as "document parsing relies on consistency in the structure of the text and any related markup language" (Loughran and McDonald 2016, p. 1192). This requirement is not always given in real-life applications, where data is often unstructured or inconsistent (over time)[6]. Other problems connected to the parsing procedure arise at the stage of tokenization or sentence-/paragraph-level segmentation. In this context, Loughran and McDonald (2016, p. 1215) provide a good introductory overview of challenges and tripwires when applying natural language processing in the financial domain. In particular, the authors highlight issues which arise from what might sound like an obvious statement, namely that"all textual methods are based on first identifying words". However, the undergoing of identifying *words* is more subtle than it appears at first sight. One needs to clarify what happens, for instance, with compound words that are connected via hyphens (e.g., well-known or up-to-date), proper nouns (e.g., the Fama and French from the Fama-French-Model), and abbreviations (e.g., FC for football club or Mr. for Mister). Moreover, the example chosen for proper nouns reveals yet another problem: how can one deal with polysemy, i.e., the fact that the surname French (of famous economist Kenneth French) looks fully equivalent to the description of nationality

[4] Most often this assumption reflects the idea that documents can be represented as so-called bags of words which allow the researcher to conduct an analysis on token-level, while at the same time this approach assumes that the ordering of words within a documents is irrelevant. I will further elaborate on this assumption in section 3.1, when I introduce the vector space model that will be the starting point for the analysis in this thesis.

[5] In most cases, the property of being machine-readable is a minimum requirement for any corpus, so this is a rather theoretical side-note.

[6] One such issue will be further explained in the description of the corpus of 10-K* filings used in this thesis (see section 5 about data collection and, specifically, footnote 49)

("to be of French origin/nationality"), yet the two words have very different meanings? This would require the researcher or the machine to learn from context, which in turn calls for usage of constructs that go above the granularity of a word-based analysis. For instance, this could be learned using n-grams (recognizing that the name French will likely tend to co-occur with words like Mr., Fama, Model, etc.; whereas the nation-based version will rather co-occur with British, German, and so on). Furthermore, conjugations and declinations can impose problems: words like calculate, calculates, calculated, calculating, etc. all refer to the same word stem, yet are treated differently just because of their suffix. As will described in section 3.1, lemmatization and stemming deal with that exact problem – in most cases by chopping off the endings and summing up those four instances as one single token (calculat).

Similarly, also an apparently trivial task like splitting a document into sentences can prove to be tricky in practice. Generally, the parsing algorithm needs to be instructed on how to treat common sentence delimiters such as periods (.), colons (;), question marks (?), or exclamation marks (!). As Loughran and McDonald (2016, p. 1216) evidence, the usage of "extensive lists, technical terminology, and other formatting complexities, makes sentence disambiguation especially challenging in accounting disclosures". An illustrative example of this challenge would be the treatment of section headers such as 2.1. Financial Statements or decimal separators as in 199.99 USD. An inaccurate parser will – based on the presence of the period sign – mistakenly split the document at these points; and "the parsing errors in this case can be extraordinary", implying that "Generic sentence parsing algorithms do not work well on financial documents" (Loughran and McDonald 2016, pp. 1216, 1217). This imprecision has crucial impact on the variable construction, implying the latter are therefore inherently measured with noise. For instance, most document readability measures are constructed using the number of sentences in a document; and thus critically depend on how accurately sentences can be identified.

2.1.1 Readability

Pioneering contributions in the stream of literature connected to readability of finance- and accounting corpora are F. Li (2008), who evidenced that firms with lower current earnings (or earnings decreases) on average tend to produce less readable as well as longer 10-K filings[7], and De Franco et al. (2015). While the former "suggests a clear correlation between the linguistic features of annual reports and firm performance" (F. Li 2008, p. 222), De Franco et al. (2015) investigate the readability of security analysts' research reports.

[7] The argument provided suggests that management has an inclination to produce longer and more complex 10-K filings so as to diffuse or dilute bad news to investors. F. Li (2008, p. 221) labels this phenomenon "management obfuscation hypothesis".

They find report readability to positively influence stock trading volume[8], with the latter being a commonly used predictor for stock return volatility and thus relating to this thesis indirectly.

In a similar setting, Hsieh et al. (2016) relate stock price reactions to the degree of readability upon publication of analyst reports. They find that the cumulative abnormal return (CAR) in a $[-1, 1]$-day window around the report issuance date increases by 58 basis points if the readability of the report increases by one standard deviation. A particularly important side note for this thesis is the statistical significance that Hsieh et al. (2016) as well as De Franco et al. (2015) find for one of their control variables, namely the narrative tone that analysts use – thus indicating that not only readability but also textual sentiment matters for the market in determining price and volume for equity instruments[9].

In the area of volatility prediction based on readability of financial documents, the most relevant paper for this thesis is Loughran and McDonald (2014). They evidence that volatility significantly decreases with the degree of readability of 10-K reports. Suggesting another, more indirect, channel of cause and effect in this context, Lehavy et al. (2011) have shown that less readable 10-K filings tend to be connected with increased analyst coverage[10]. Referring to the findings of Schutte and Unlu (2007), who evidence that analyst coverage initiation helps investors to better distinguish "true" firm-specific information from pure noise signals, it can be inferred that the fact that an analyst tracks the company *per se* has a negative impact on firm-specific volatility. This interpretation undermines the findings of Loughran and McDonald (2014), namely that more readability in 10-K's (be it directly or via increased analyst coverage) reduces the stock price volatility of the companies submitting the filings.

Regarding the **measurement** of readability, different metrics have evolved in the literature, with the most commonly used being the Gunning-Fog-Index as well as the Flesch-Kincaid-Grade-Level-Formula and the Flesch Reading-Ease Score (FRES)[11]; nevertheless, Loughran and McDonald (2014, p. 1649) argue that these measures, because of their strong focus on sentence length and word complexity, are not suitable in the context of business-related text. Moreover, they do not adequately account for differences in background knowledge between the targeted audiences. In fact, based on the Fog measure, very common words

[8] A potential explanation for this reads as follows: if a report, ceteris paribus, classifies as "more readable", the information conveyed in the document will be perceived as more precise information and thus trigger stronger trading signals for investors (De Franco et al. 2015).

[9] De Franco et al. (2015) additionally include an interaction term between readability and tone, which they find to be positive and significant. This suggests that these two variables reinforce each other.

[10] The authors provide a potential explanation for that phenomenon: a 10-K that is relatively difficult to read has too high processing costs for a single analyst to cover them. Thus, naturally, in order to meet investor's information demand, more analysts will start to track the stock and exert collective effort.

[11] All of these measures effectively are a linear combination of two components, namely the average length of sentences as well as proportion of "complex" words used, while complexity of words is determined by syllabication. A more detailed definition of these metrics is provided in appendix **??**.

like company, management, or financial would be considered to be "complex" words – yet they are still very likely to be understood by the average investor. Furthermore, all of the common readability measures require some form of text parsing and are thus exposed to typical inaccuracies and subjectivities connected to such procedures. In the work of Loughran and McDonald (2014), the "underlying" textual source were 10-K filings with the SEC. For these documents the authors suggest an alternative that is a much simpler, less error-prone and more reproducible metric, yet is still correlated to "conventional" measures like the Fog Index and equally able to capture readability: the natural logarithm of gross file size (measured in megabytes) of the 10-K filing[12].

2.1.2 Sentiment Analysis

As Loughran and McDonald (2016) indicate, the most simple approach to a textual analysis with financial corpora is to search the documents for appearance of certain keywords (or keyword phrases) of interest (for instance, they cite phrases like ethic or corporate social responsibility, whose frequency can be used when one attempts to quantify corporate governance metrics or the probability of lawsuits). A natural extension of this methodology involves to not look for specific phrases only, but rather use a dictionary of words[13], all of which carry similar characteristics. In the large majority of cases these word lists are chosen so as to share common sentiment and thus relate directly to the technique of sentiment analysis.

The frequency of sentiment words is then often aggregated into a scalar, constructing what is often referred to as "sentiment index" or "sentiment score"[14]. Apart from the usability as a variable in many research applications, sentiment indices are common also in practice, as the following quote undermines:

> Gauging sentiment with an index makes it possible for machines to analyse the information; it can be converted, aggregated, and compared. And the index can be used with statistical analysis to build prediction models. Obviously, the difficulties come in the details of assigning the sentiment index. At the core of the process is a lexicon that lists words or phrases that represent a certain kind of sentiment and, importantly, reflect the specific context in which the text appears. (Kremer et al. (2013))

[12] One should note at this stage that this approach requires the researcher to account for firm complexity (most likely using a proxy like firm size), as longer reports could be simply due to the complexity of the fundamental business (Loughran and McDonald 2016).

[13] Throughout this thesis the words lexicon, dictionary, or (sentiment) word list will be used interchangeably; although from a linguistic and literary point of view they are not equivalent.

[14] Note that the strong focus on keyword counting in sentiment analysis has lead the discipline to be dominated by vector space models (or, equivalently, term-document-matrices) within the bag-of-words framework (see section 3.1 where these concepts will be expanded in the process of introducing the research design applied in this thesis.

The last part of the definition highlights a critical fact that distinguishes sentiment analysis in finance and accounting from other disciplines: while early studies used "generic" lexica[15] that were developed from (and thus suitable for) texts in psychology and sociology, Loughran and McDonald (2011) were the first to recognize that those lists are not applicable to documents in the business domain. In fact, using 10-K filings, the authors find that almost three quarters of allegedly negative words in the Harvard IV-4 list do not appear to be negative in business context[16]. Therefore, Loughran and McDonald (2011) created finance-specific sentiment word lists in the following categories: negative, positive, uncertainty, constraining, litigious, strong/modest/weak modal. These lists (especially the negative lexicon) have become the "gold standard" in financial sentiment analysis and are also the backbone of this thesis. Henceforth, I will label these word lists using the acronym LM dictionary (or lexicon/word list).

Turning from this methodological introduction to relevant literature contributions, a good starting point for dictionary-based sentiment extraction is the paper of Yukselturk and Tucker (2015). They calculate thematic net sentiment scores for a UK-based sample of analyst research reports by classifying (80 to 180) keywords contained in six theme-buckets[17] into positive and negative keywords, while the classification task was performed using both Harvard IV-4 and LM dictionaries. The main finding was that theme-related (net) sentiment affects the two key outputs (recommendations and target prices) of the research analyst. A. H. Huang et al. (2014) extended this stream of research regarding textual contents of analyst reports. They quantified sentiment of sentences contained in analyst reports using a Naive Bayes machine-learning algorithm. After controlling for a simultaneous change in the quantitative summary metrics (i.e., price target, stock recommendation and EPS forecast), they show that the textual opinion embedded in the respective analyst report helps to explain abnormal stock returns subsequent to the release of the report. Moreover, they evidence that the narrative content of analyst reports also carries "predictive value for future earnings growth in the subsequent five years" (A. H. Huang et al. 2014, p. 2151).

In the same category of research falls also an important paper of Tetlock (2007). He mined a popular column in the *Wall Street Journal* in order to gauge market sentiment; in particular, he constructed a pessimism index using the Harvard GI lexicon, which was shown to be predictive for lower subsequent returns and higher volatility.

Another pioneering paper in this area of research, which due to its nexus to volatility prediction is also relevant for this work, is Antweiler and Frank (2004). They analyzed

[15] The most prominent choice for word classification among the heap of lexica available was the Harvard General Inquirer (GI) (in particular, the Harvard IV-4 dictionary, see `http://www.wjh.harvard.edu/~inquirer/homecat.htm`).

[16] For instance, such words are tax, cost, capital, board, liability, foreign, and vice, which appear very often in 10-K filings, yet are very likely not negatively connoted (Loughran and McDonald 2011, p. 36).

[17] The six themes are the following: macroeconomic and regulatory environment, industry and market environment, growth, management and strategy, financial performance, financial position.

more than 1.5 million messages from message boards about 45 companies; using the Naive Bayes approach to classify messages to either buy, hold, or sell category. The single classified messages were then aggregated into a bullishness as well as an agreement index, which proved to be successful in predicting both trading volume and volatility, but unsuccessful in explaining stock returns. Regarding volatility, only the bullishness seems to carry explanatory power while the correlation is positive (i.e., more bullish messages today imply greater volatility tomorrow), while agreement between message posters does not help to predict realized volatility.

Most importantly, however, this thesis builds upon the work in Kogan et al. (2009), which was extended in Tsai and Wang (2012), Tsai and Wang (2013), Tsai and Wang (2014), Tsai, Wang, and Chien (2016), and Wang et al. (2013) and Tsai and Wang (2016); all of which essentially show that textual sentiment of 10-K's helps to explain stock return volatility in the quarter/year after the filing (either in a ranking or regression framework). These results were yet further refined in Rekabsaz et al. (2017), who used a more elaborate word weighting scheme on the LM sentiment lists and also investigated on potential industry-specific patterns.

2.1.3 Text Analysis Using 10-K Filings Corpora

As it is evident from the literature review up to this point, corporate filings as a mandatorily disclosed document provide a rich source of potential research questions in the textual analytics field and are very popular within the finance and accounting domain. In this context, Das (2014, p. 223) points out[18]:

> Whereas much of financial text analysis uses messages posted to finance boards like Yahoo!, Motley Fool, or to blogs such as Twitter, Facebook, etc., other textual analysis has been applied to company reports and filings. In the former case, analysis is usually of a time series nature, whereas in the latter, text analysis is undertaken across companies in a cross-section.

Regarding the heap of corporate filings available, Das (2014, p. 227) further highlights that most of the literature is focussing on 10-K filings because these are considered to be the most informative document for (potential) investors[19]. 10-K reports are mandatory

[18] In that regard, Kearney and Liu (2014, p. 3) highlight that while corporate disclosures provide "a natural source of textual sentiment for researchers insofar as they are official releases", their main disadvantage "is the low frequency of the data, because firms usually make these disclosures on a quarterly or annual basis."

[19] I follow this stream of literature and use a broad corpus of 10-K*s (refer to footnote 2 for a description of variants of the "standard" 10-K). Moreover, to avoid permanent repetition I will use the terms 10-K, filing, or (annual) report synonymously, although the SEC highlights that "10-K typically includes more detailed information than the annual report to shareholders." (see `https://www.sec.gov/fast-answers/answersreada10khtm.html`, visited on 05/27/2018)

disclosure documents that are to be filed annually with the SEC. As it is outlined on their web-page[20], the report's contents contain an important source of information and decision-making material for investors: "Among other things, the 10-K offers a detailed picture of a company's business, the risks it faces, and the operating and financial results for the fiscal year. Company management also discusses its perspective on the business results and what is driving them." As Pulliza (2015) points out and what is of direct importance for all language analytics studies using this corpus, 10-K filings shall be conform with the SEC Plain English Rule, which was established with the target that communication from the management towards the investor community is written in plain and understandable terms without an overload of complexity.

One should note that this part of the literature review reflects only "related" contributions, i.e., publications with a focus on natural language processing and, more specifically, sentiment analysis. The (much broader) stream of literature that deals with the "quantitative" outputs of (mandatory) corporate disclosure and the information content as well as the market reactions of such disclosures, is not part of this review. This, for instance, encompasses a large range of publications in the field of accounting policies (accounting standards, international comparisons, etc.) and manipulation methods (with contributions related to (abnormal) accruals mainly) as well as mergers and acquisitions including the subsequent purchase price allocations (with contributions related to intangible assets, goodwill, etc.) that are disclosed in annual filings. In fact, I view the combination of such "hard" information content with the textual tone as a much desired extension for further research; a fact on which I will further elaborate in the conclusions of this thesis (cf. section 8).

One of the first questions to ask when analyzing 10-K filings is whether to text mine the whole document or rather target a specific section in it. The most famous of them, which is often said to be the most informative part and carry the most forward-looking statements, is Item 7 of the 10-K filing, which is referred to as *Management Discussion and Analysis* (MD&A). For instance, this section is the subject of study in widely cited papers like Kogan et al. (2009), F. Li (2010), Feldman et al. (2010), Tsai and Wang (2012), Wang et al. (2013), Tsai and Wang (2013), Tsai, Wang, and Chien (2016), and Tsai and Wang (2016).

Another common choice is *Item 1A - Risk Factors*, "which contains information about the most significant risks for the company" (Rekabsaz et al. 2017, p. 1712), which is used, among others, in K.-W. Huang and Z. Li (2008) and Rekabsaz et al. (2017). A good overview of all sections within a 10-K filing is provided on the SEC web-page[21].

However, also section-based analysis comes with obstacles: as Loughran and McDonald (2016) evidence, parsing algorithms struggle with the fact that some 10-K submissions are

[20] See `https://www.sec.gov/fast-answers/answersreada10khtm.html`, accessed on 07/31/2018.
[21] See `https://www.sec.gov/fast-answers/answersreada10khtm.html`, accessed on 7/31/2018

unstructured (especially before 2002), some sections are mislabelled (e.g., MD&A is falsely tagged as Item 6 instead of 7), or that companies place content across different parts of the document and cross-reference across the filing using footnotes. Regarding the latter, Heidari and Felden (2015) presented a promising idea, that might be an interesting extension to 10-K corpus based research: they tried to categorize income tax footnotes within 10-K and 10-Q's into pre-defined buckets and proved that machine-based classifiers can achieve the same task with relatively high accuracy, allowing the researcher to extract content from an "unstructured" part of the filing without the necessity of reading the footnotes manually. This was extended upon by Amel-Zadeh and Faasse (2016) and Thinggaard et al. (2015), who compared MD&A content with footnote disclosure, the latter doing so for the Danish market and therefore belonging to the absolute minority of publications which use non-US based data.

The most influential papers who focus on the 10-K as single, homogeneous document are Kothari et al. (2009), Lehavy et al. (2011), Loughran and McDonald (2011), Loughran and McDonald (2014), and Loughran and McDonald (2016), while good introductory overviews about 10-K language processing (and, more generally, textual analysis in finance and accounting) are provided by Qiu (2007), Pulliza (2015), and Loughran and McDonald (2016).

3 Term Weighting and Calculation of Sentiment Scores

This section outlines the first part of my research framework, namely the sentiment extraction process from the 10-K* filings. Starting with simple term counts and conventional weighting schemes, in the course of this section, I will also introduce the "volatility-impact-based" term weighting building on Jegadeesh and Wu (2013); in this context, the volatility measure applied in this thesis will be explained. Moreover, the section also describes the aggregation of (weighted) sentiment frequencies into sentiment scores.

3.1 Term Counts, Document-Term-Matrix, and Vector Space Model

When considering sentiment analysis, a common starting point is the representation of the textual corpus $\mathcal{D}$ in a so-called document-term-matrix, commonly abbreviated by DTM[22]. As indicated, a corpus, denoted by $\mathcal{D}$, is simply a collection of N different documents which the researcher seeks to analyze, i.e., $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$. Thus, in a DTM each of the rows of the DTM is representing one document d_i, $i = 1, \ldots, N$, from corpus $\mathcal{D}$. In the case of this thesis, this corresponds to one 10-K* filing per row. Similarly, each column stands for one word *type* in the available vocabulary[23]. The vocabulary in turn can be either formed in a "naive" way, i.e., include all terms that appear in the full corpus at least once (in other words, in at least one document) or, alternatively, contain only pre-selected words from a given lexicon. In this thesis, I will use eight different lexica produced by Loughran and McDonald (2011). The main reasons for this are indicated in the enumerated list in the subsequent section 3.2, in which I discuss crucial elements in determining the size of the vocabulary. Henceforth I will indicate the corresponding LM-lexicon to which a word belongs by using sub- and superscript k, with $k \in \{N, P, U, L, C, SM, MM, WM\}$, which stand for negative, positive, uncertainty, litigious, constraining, strong modal, modest modal, and weak modal, respectively[24]. Details on the LM word lists and their composition can be found in Appendix A.1. The dictionary of category k, denoted by $\mathcal{J}_k$, has length

[22] One can, alternatively, also transpose the matrix and represent the corpus as a term-document-matrix, TDM.

[23] Althoug henceforth the term "word" will be used synonymously to the phrase "word type", it is at this stage useful to briefly present an important concept in text analysis: the distinction of word **type** and **token**. While types represent the number of *distinct* words within a corpus, tokens are the total number of all words (Jurafsky and Martin 2017). Very common phrases like `and`, which appear very often in the corpus, imply that there are many tokens for this word but in fact only one type. This also highlights the usefulness of a matrix representation, in which one can take account for multiple occurrences of a word type by displaying its *count*.

[24] All dictionaries are provided by Loughran and McDonald (2011) (`https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary`). The constraining as well as weak and modest modal word lists are used in the linguistic analysis only, but are unrelated to the research hypotheses presented in this thesis (see chapter 4).

J_k, i.e., consists of J_k different words labelled $v_{1,k}, v_{2,k}, \ldots, v_{J_k,k}$. Thus, one can write the resulting eight DTM's in the following general notation:

$$DTM_k = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_N \end{array} \begin{array}{c} \begin{array}{cccc} v_{1,k} & v_{2,k} & \ldots & v_{J_k,k} \end{array} \\ \begin{pmatrix} tf^k_{1,1} & tf^k_{1,2} & \ldots & tf^k_{1,J_k} \\ tf^k_{2,1} & tf^k_{2,2} & \ldots & tf^k_{2,J_k} \\ \vdots & \vdots & \ddots & \vdots \\ tf^k_{N,1} & tf^k_{N,2} & \ldots & tf^k_{N,J_k} \end{pmatrix} \end{array} \tag{3.1}$$

As said, DTM_k has dimension $N \times J_k$, where the row dimension, N, represents the number of documents (i.e., 10-K* filings) in the corpus. The number of columns, J_k, denotes the number of sentiment words $v_{j,k}$, $j = 1, 2, \ldots, J_k$, available in the LM dictionary of category k[25]. Finally, and most importantly, each coefficient of the DTM, $tf^k_{i,j}$, with $i = 1, \ldots, N$ and $j = 1, \ldots, J_k$, denotes the number of occurrences of word type j from LM-list k in document i (so-called raw "term frequency" (TF))[26].

Inspecting the rows of the DTM, one can see that each of the N documents can be represented as an J_k-dimensional vector of word counts. Thus, this formulation is often referred to as "Vector Space Model" and is especially common in information retrieval applications, as it intuitively formalizes the idea that two similar documents will tend to contain comparable sets of words. Hence, their word count vectors should also be similar[27]. However, it is of crucial importance to note an underlying assumption that is

[25] One further remark can be made regarding the lexicon size: if one were to create one "global" DTM with all J (stemmed) word types (coming from all eight available LM dictionaries), the dimensionality would not simply add up: Thus, $J = 1,574 < J_N + J_P + J_U + J_L + J_C + J_{SM} + J_{MM} + J_{WM} = 1,714$. The inequality comes from the fact that the sets $\mathcal{J}_k$ are not disjoint (i.e., several words occur in more than one list). For instance, the uncertainty and litigious sub-lexicon share some common words with the negative word list. Details regarding the LM lexica as well as their overlap are outlined in Appendix A.1.

[26] There are also "simpler" variants of the DTM in which the coefficients do not represent the absolute term frequencies, but rather an indicator variable, i.e., a Boolean equalling one if the term appears (at least once) in the document, and zero else. This is often referred to as term-document **incidence** matrix (Manning et al. 2008, p. 4).

[27] The general task in information retrieval would be the detection of a set of documents, each represented by a J_k-dimensional document vector $\mathbf{D}$, that are most similar to a given set of search-/keywords, represented by a J_k-dimensional query vector $\mathbf{q}$. The similarity between two vectors in most applications is measured by so-called **cosine similarity**, i.e., the cosine of the angle that is spanned between the two vectors. Mathematically, one takes use of the notion of dot product and defines:

$$\cos(\theta) = \frac{\mathbf{D} \cdot \mathbf{q}}{\|\mathbf{D}\| \cdot \|\mathbf{q}\|} = \frac{\sum_{j=1}^{J_k} D_j q_j}{\sqrt{\sum_{j=1}^{J_k} D_j^2} \sqrt{\sum_{j=1}^{J_k} q_j^2}},$$

where $\|\mathbf{z}\|$ is called the **Euclidian norm** (commonly referred to as "length") of the vector $\mathbf{z}$ and is used "normalize" vectors of differing lengths. Generally, $\cos(\theta) = -1$ for vectors pointing in opposite directions, $\cos(\theta) = 0$ for orthogonal vectors, and $\cos(\theta) = 1$ for vectors pointing in the same direction. However, as word counts are always non-negative ($tf^k_{i,j} \geq 0$), cosine similarity in text classification problems always lies between zero and unity (Jurafsky and Martin 2017, p. 280).

implicitly made, if one seeks to represent the corpus in a DTM: if all documents can be displayed as vectors of term counts, the ordering of the words within the document is lost. This is commonly referred to as **bag-of-words** assumption; specifically, one represents each document as "an unordered set of words with their position ignored, keeping only their frequency in the document" (Jurafsky and Martin 2017, p. 76). Therefore, the two sentences Martin is faster than Tom and Tom is faster than Martin have two equivalent bag-of-words vector representations, although their meaning is oppositional.

Besides the implicit bag-of-word assumption, another common feature of vector spaces, also due to the often very high dimensionality of the DTM, is the so-called *sparsity*, which refers to the fact that most of the elements within the vectors will be zero (Jurafsky and Martin 2017, p. 252). This circumstance stems from the following observation: for a word j to be included in the DTM, it is sufficient that the j-th column sum of the DTM is strictly greater than zero; however, this condition does not prevent many row sums to be zero. Thus, it is enough for the word to appear in only one of the documents to be part of the DTM, thereby creating a lot of zero entries for all other documents where it does **not** appear in. Especially (but not exclusively) in text *classification* tasks the sparsity of the DTM can be problematic[28]. Due to this reason, researchers often apply so-called *smoothing* (or discounting) algorithms to the DTM. The most common choices are Laplace and "add-k" smoothing. The former simply replaces $tf^k_{i,j}$ in the DTM by $tf^k_{i,j} + 1$, i.e., increases all term counts by one. Similarly, the "add-k" smoother fills the cells of the TDM with $tf^k_{i,j} + k$ with $k \in (0, 1)$ (Jurafsky and Martin 2017, pp. 47 sqq.).

3.2 Dimensionality of the Document-Term-Matrix

In discussing the DTM in general (and also with respect to the research specification in this thesis), a few other aspects are crucial, as they significantly affect the row dimension (J_k) of the matrix:

(1) **Choice of the vocabulary.** As already indicated, in this thesis the row dimension of DTM_k will be equal to the number of all unique, stemmed words in sentiment list k. The other common choice is to let the row dimension be equal to **all** words that are available in the entire (training) corpus. This number will, in virtually all practical

[28] The problem with sparsity arises most often in cases where the training (in-sample) data contains a zero count for a particular word *and* class, whereas the term occurs in the test set (out-of-sample) for the first time in that specific class. Many (probabilistic) classifiers will at least in some form be based on the prior (conditional) probability (i.e., observed frequency) of the class given the term, which will be zero; thus underestimating the probability of occurrence. Note that this is not to be confused with the case where terms are non-existent in either class (so-called "unknown" or "out-of-vocabulary" (OOV) words), which in most cases are either fully omitted from the analysis or, alternatively, are added to the word list with an additional suffix tag like `UNK` (for unknown) (Jurafsky and Martin 2017, pp. 46, 79).

cases, massively exceed the number of LM-sentiment words[29]. Consequently, reducing the analysis to a small subset of words in the corpus that one deems to carry "sentiment influence" might seem like a severe restriction. However, Tsai and Wang (2016) have extensively compared vector spaces using "original" versus "sentiment-limited" dimension and have shown that the latter perform (at least) equally good[30]. Regarding the choice of full versus pre-selected word list, also Loughran and McDonald (2016, p. 1215) speak in favour of using pre-defined lexica by pointing out that using the full list of corpus words "can produce a relatively long list that is substantially and meaningfully shortened by selecting only those tokens that map into a list of words." This is mainly due to the fact that many words in the original, full word-list are "meaningless" for sentiment analysis purposes. At the same time, this dimensionality reduction critically eases computational effort and is therefore also chosen for the project in hand.

(2) **Stop words.** Stop words are very frequent words like the, a, of, or, in, that, to, and, with or for, which have very high word counts, yet are semantically "meaningless" (Gries 2009; Jurafsky and Martin 2017). Very often one would therefore exclude such words from the DTM. The question of stop word exclusion partly relates to vocabulary choice described above: calculating sentiment scores based on pre-defined sentiment lists instead of the original full-corpus vocabulary, does eliminate the necessity of defining and removing stop words from the corpus, as most common stop words do **not** appear in either of the LM sentiment lists[31]. However, the choice whether to in- or exclude them in the DTM might be very important for *other* purposes. For instance, measuring document length by some function of the number of types or tokens, will produce very different results depending on whether stop words are considered or not. Moreover, many measures of document readability are based on syllabification. Since stop words are often monosyllabic, the in-/exclusion of such words will drastically influence the readability measure. As a final example, one might also think of a potential overlooking of negation, as words like not or no are included in some stop word lists.

(3) **Lemmatization and stemming.** The process of lemmatization refers to "the task of determining that two words have the same root, despite their surface differences", whereas stemming "refers to a simpler version of lemmatization in which we mainly

[29] Jurafsky and Martin (2017, p. 274) report that this number usually is somewhere between 10,000 and 50,000 words and thus much larger than the length of any meaningful sentiment lexicon.

[30] It shall be highlighted that this methodology remains tested solely under the bag of words assumption. Moreover, many papers in existent literature do choose the same approach; however, many expand the LM-dictionary by also including the top-related terms in the total corpus-wordlist (determined by the usage of cosine similarity of such words to those in the respective LM list). Examples of this technique include Tsai and Wang (2014) and Rekabsaz et al. (2017).

[31] In this context, Loughran and McDonald (2016, p. 1206) state: "Given that most business applications of textual analysis focus on using word counts from sentiment categories, the elimination or special treatment of stop words is typically not necessary."

just strip the suffixes from the end of the word" (Jurafsky and Martin 2017, pp. 11, 25). An example for lemmatization is to replace words like am, are, is, were, or was with their common root (to) be. The impact on the DTM is evident; in this case the dimension is heavily affected (instead of five words, only a single lemma would be left appearing in the DTM). Similarly, an example for stemming replaces words such as computer, compute, computed, computational, or computationally with their shared stem comput and thereby reduces word-dimensionality in the DTM as well[32]. The choice of an appropriate lemmatizer or stemming algorithm is of crucial importance, as both procedures can be error-prone: for example, a stemming algorithm might stem the word presentation to present, thereby mingling it with the noun present (like gift), which then in turn itself might be confused with a form of present tense/time. The most widely used stemming algorithm, which is also applied to the LM word lists used in this thesis project, is the *Porter Stemmer*, which is very simple and efficient; yet one should note that, as Jurafsky and Martin (2017, pp. 25, 26) illustrate, it is still exposed to make mistakes in terms of both over- and under-generalization.

(4) **Negation tagging.** Negation tagging refers to the process of identifying semantical negations by adding pre- or suffixes to words in case that they co-occur with typical negation phrases like no, not or never. For illustration purposes, one can consider the following two sentences: The company did **not** report a loss for three consecutive years versus The company did report a loss for three consecutive years. The inclusion of the bold-faced word not completely reverses the sentiment extracted from two otherwise identical sentences. Negation tagging would transform the first sentence to either The company did_not report a loss for three consecutive years, The company did report_not a loss for three consecutive years or even The company did report_not a_not loss_not for_not three_not consecutive_not years_not, where in the last version the negation tag is suffixed to each word until the next punctuation mark appears (Jurafsky and Martin 2017, p. 81).

In the context of negation tagging, the most critical issue that needs to be addressed is the choice of the "negation window" around the word under consideration, i.e., the size of the n-gram[33] around the current word in which one looks for negation phrases. For example, in the sentence The weather today was not rainy but rather sunny applying a negation tag also to the word sunny would alter the meaning of the sentence twice. However, when inspecting the word sunny, the choice of the n-gram obviously determines "how far around" this specific word one would search for negation phrases. In this case

[32] Jegadeesh and Wu (2013) manually inflected the LM dictionaries and thereby reduced the number of sentiment words by around one third, namely from 353 (2337) to 122 (716) for positive (negative) words.

[33] The notation of a **n-gram** simply refers to a sequence of n words. Similarly, the notation of unigram ("single word"), bigram ("word pairs"), or trigram ("word triplets") refers to the most common choices of $n = 1$, $n = 2$, and $n = 3$, respectively.

the critical question is whether n is large enough so as to capture the fourth word preceding sunny (which would be the negator, not). Additional complexity arises due to the fact that this n-gram can be designed in three ways, i.e., considering preceding and/or subsequent words.

Yet even this trivial example shows that there is no generally acceptable optimal solution regarding the choice of the appropriate n-gram, as the "searching-for-negation-window" shifts along as one continues to read. Particularly, in this example one would likely be interested to tag the word rainy transforming it to rainy_not. This implies that $n \geq 1$ and inclusion of preceding words is required. In order to keep unchanged the words but and rather – assuming they were not classified as stop words in the first place and therefore removed – one must set $n = 1$ or $n = 2$, respectively. The word sunny would remain untagged even if one considers a trigram of preceding (!) words[34].

(5) **Numerical characters.** It is also important to clarify how to deal with figures that occur in the text body of the document. Some researchers and practitioners exclusively focus their analysis on the textual contents within the corpus and thus decide to delete sequences of numerical characters. Others, however, choose to replace them by a common representative symbol, like the hash sign (#). Yet, even those apparently simple parsing procedures have their caveats. One might consider, for instance, a price tag of 99.99 USD appearing in one of the documents. If the parsing algorithm mistakes the decimal separator (.) to be a sentence separator, this string would be replaced by #.# USD, failing to recognize that it is in fact a single number. Therefore, the "word" # would falsely appear twice (instead of once) in the DTM. Similarly, contingent on how the parser deals with the / sign, a date string like 05/23/2009 might be transformed into three hashes, although it really refers to one single date. Common solutions to such problems include the usage of so-called *regular expressions* (RegEx), which allow to search for detailed patterns in strings[35]. Moreover, another problem in replacing numbers with the hash sign includes that the numerical count might be upward biased simply due to the presence of page numbers in the footer, consecutively numbered section headers, or the usage of enumerated instead of bullet-pointed lists.

To address these issues, I will apply the following parsing procedure to the 10-K* sample used in this thesis:

[34] Note that in this case it is pointed out that the trigram covers only leading and no lagging words, as it is not clear what follows *after* the word sunny. For instance, if the next sentence reads On no other days it was sunny anymore, and the trigram is symmetrically around sunny, then the word would be again falsely tagged due to the lagging negator in the next sentence.

[35] In this case, in order to correctly classify a date, one would look for strings with the following pattern: one or two digits, followed by /, followed by one or two digits, followed by /, followed by four digits. It is easy to verify how such a RegEx would expose the researcher to the risk that dates might as well be formatted as YYYY/MM/DD or could, alternatively, be separated by the symbols . or - instead.

(1) As indicated, I will extract one DTM_k for each LM-lexicon k, thereby capturing 1,562 words out of the full (stemmed) LM word list with 1,574 unique terms (see Appendix A.1 for details).

(2) Using only words in the respective LM-category k, only six stop words appear in any of the DTM's: against, could, further, ought, should, would[36].

(3) The words in the LM lists were stemmed using the Porter stemmer embedded in the `tm` package in `R`. The full lexicon of stemmed words is provided in Appendix A.1.

(4) Due to the complexity attached to negation tagging, missing evidence from the literature regarding the size of the n-gram window as well as non-existence of negation word lists for financial corpora, no negation tagging was applied in this thesis. As Loughran and McDonald (2016, p. 1217) point out, negation typically occurs with positive words, as the management of the filing company seeks to disguise bad messages using positive language; conversely, the authors state that "negative words seem unambiguous – rarely does management negate a negative word to make a positive statement", thereby inducing enhanced caution when interpreting results based on positive sentiment.

(5) As the focus of this work is to investigate the impact of textual contents within corporate filings, numerical strings are removed from the corpus. Moreover, any variable constructed using counts of numeric characters would likely be subject to measurement error, as parsing can prove difficult for finance corpora. However, partly acknowledging the importance of messages conveyed by numbers, ratios, and statistics, I will capture the degree of "numerical information" by the number of occurrence of frequent companions as percent, dollar, or Euro (for details see section 4.2 on variable construction).

3.3 Transforming Term Counts to Term Weights

Sentiment analysis literature, again borrowing from concepts in information retrieval, quickly realized that *raw* term counts in the DTM suffer from significant drawbacks.

Firstly, term frequencies scale with document length. This implies that longer documents will in most cases also have higher term counts[37]. Therefore, it is common to opt for the use of **relative** term frequencies instead of absolute term frequencies, implying that

[36] The small overlap of the stop word list and the eight LM dictionaries is based on the stop word list provided by the `tm` package available for the programming language `R`, in which this research project was conducted. However, using other common stop word lists delivers highly similar results.

[37] Manning et al. (2008, p. 127) provide an intuitive example for this fact, namely when a document d_i is simply cloned and combined with itself. Raw term frequencies in this case double, although the "informational content" of the document after this manipulation is the same as before.

each count is scaled by document length ($L_{i,k} = \sum_{j=1}^{J_k} tf_{i,j}^k$)[38], thereby obtaining $rf_{i,j}^k = tf_{i,j}^k / L_{i,k} = tf_{i,j}^k / \sum_{j=1}^{J_k} tf_{i,j}^k$. This approach was, for instance, used in Tsai and Wang (2014) and Tsai and Wang (2016). In some occasions, also a simple Boolean indicating occurrence versus non-occurrence of the term – without considering the "degree" of occurrence – might mitigate the length-scaling problem of raw term counts (see footnote 26 on term-incidence-matrices).

Secondly, a comparison that is based solely on raw frequency implies that a word $v_{j_1,k}$ with $tf_{i,j_1}^k = 10$ is *ten times* more important than a word $v_{j_2,k}$ with $tf_{i,j_2}^k = 1$. In fact, it is often argued that actually the more seldom / unique words are "more important", as they help the reader to distinguish one specific document, in which they appear, from the other files in the corpus. For this reason, many studies apply a smoothing correction to raw term frequencies and use logarithms as transformation. The modified counts are often called "weighted" frequencies:

$$wf_{i,j}^k = \begin{cases} 1 + \ln(tf_{i,j}^k) & \text{if } tf_{i,j}^k > 0 \\ 0 & \text{else} \end{cases} \tag{3.2}$$

Sometimes, also $wf_{i,j}^k = \ln(1+tf_{i,j}^k)$ is used so as to smooth the positive term counts (e.g., in Rekabsaz et al. (2017)). I will refer to these weighting schemes as `WF_1PLOG` and `WF_LOG1P`, respectively, and I will use them as benchmark weighting schemes in robustness checks.

Yet another alternative and commonly used weighting scheme is to normalize all term counts by the count of the most common word in the respective document, i.e., the largest term frequency observed in the respective document d_i. Denoting the latter by $tf_{i,j_{max}}^k$, one can scale as follows:

$$maxtf_{v;d} = a + (1-a)\frac{tf_{i,j}^k}{tf_{i,j_{max}}^k}, \tag{3.3}$$

where a is a smoothing parameter that is usually set to .4 by convention (Manning et al. 2008, p. 127). In other words, normalized term frequencies are obtained by scaling all elements in the DTM with the maximum term frequency of each respective row of the DTM in which they appear (Manning et al. 2008, pp. 126-127). In robustness checks, I will label this weighting approach `TFMAX`.

The third and most critical issue of applying raw/relative/weighted term frequencies stems from the following fact: as they are based on term counts *within* a single document i, two sentiment words with equal frequency are considered equally important, regardless

[38] One could, alternatively, also define document length in many other ways: For instance, one could simply count all number of tokens or types in the document. Similarly, one could also use the count of **all** sentiment words (L_i) instead of using category k-specific lengths. However, I find this approach more suitable for various reasons: Firstly, it assures that each row in each DTM_k sums to unity. Secondly, using $L_{i,k}$ expresses the relative frequency within a group of similar words and thereby mitigates a potential dilution of the relative frequency in category k_1 due to high occurrences of words from category k_2. Thirdly, as I will compute sentiment-category specific scores as well, I prefer to operate with a category-specific length measure.

of how often they appear in **other** documents. This notion is critical, as one can illustrate with a brief example: the two negative LM words defer and cyberattack might have term counts of, say, 20 each in a given document d_m. However, assume that defer appears in *all* other documents d_i (with $i = 1, 2, \ldots, N$) of the corpus as well (for instance, due to usage of common phrases like deferred tax asset/liability); while cyberattack appears exclusively in d_m. It is obvious that the term cyberattack will have much more "explanatory power" in expressing a sentiment in that specific document d_m, although it has the exact same term frequency as defer. In contrast, the latter offers no "discriminatory" power whatsoever in order to distinguish document d_m from all the others in the corpus. This example undermines why this issue was first addressed in information retrieval tasks, where distinction of documents is crucial. However, also sentiment analysis commonly borrows from the solution that was proposed to solve this issue, which is presented in the following subsection.

3.3.1 Term Weighting Schemes Using Inverse Document Frequency

The most common choice to account for the differing occurrence of words within the whole corpus is to scale (raw or weighted) term counts with a measure called **inverse document frequency** (IDF) and create a combined measure called TF-IDF. Inverse document frequency attaches high weights to words that appear in few documents in the corpus, and low weights for words that are very common across documents in the collection. Formally,

$$idf_{j,k} = \ln\left(\frac{N}{df_{j,k}}\right), \tag{3.4}$$

where N, as usual, stands for the number of documents in the corpus and $df_{j,k}$ denotes the number of documents in which term $v_{j,k}$ appears at least once. In other words, $df_{j,k}$ corresponds to the number of non-zero entries for each **column** of the DTM. Note how, therefore, $idf_{j,k}$ has no document subscript i: being based on the columns of the DTM only, it is not a document-specific but rather a word-to-corpus-based measure. Moreover, and from a practical perspective, there is no consensus about which base of the logarithm is most appropriate (in both equations (3.2) and (3.4)), while most applications opt for either natural base, $\log_2$ or $\log_{10}$[39].

Using inverse document frequencies computed by equation (3.4), the "combined" TF-IDF weight for term $v_{j,k}$ in document i then becomes:

$$tfidf_{i,j}^{k} = tf_{i,j}^{k} idf_{j,k} \tag{3.5}$$

[39] In all subsequent applications in this thesis, natural logarithm will be used, while notations log and ln will be used interchangeably.

This weighting scheme will accordingly be named `TFIDF`.

If one instead opts to choose either form of the log-normalized variants for the term frequencies, $wf_{i,j}^k$, the IDF-method accordingly defines:

$$wfidf_{i,j}^k = wf_{i,j}^k idf_{j,k} \tag{3.6}$$

Depending on the definition of $wf_{i,j}^k$ (using either $(1 + \ln(x))$ or $\ln(1 + x)$), this results in weighting schemes `WFIDF_1PLOG` and `WFIDF_LOG1P`, respectively. Similarly, one obtains `RFIDF` by:

$$rfidf_{i,j}^k = rf_{i,j}^k idf_{j,k} \tag{3.7}$$

As one can see from equations (3.5), (3.6), and (3.7), the IDF-based weighting schemes favour those words that are frequent in one document i but at the same time are rare in all other documents of the corpus. Hence, such weighting methods solve the problem described in the introductory example in section 3.3: by attributing to the "common word" defer a lower IDF value than to the "one-document-only" term cyberattack, the latter will, for equal term counts, have a higher IDF weight. For this reason, these weighting schemes are also a suitable method to detect and "filter out" stop words: it can be easily verified that for terms (such as and or a), which appear in *all* documents, one obtains $df_{j,k} = N$ and therefore $tfidf_{i,j}^k = wfidf_{i,j}^k = rfidf_{i,j}^k = 0$. Thus, even if one fails to exclude the very common and "meaningless" words, an IDF-based scheme would weight those terms close to zero and "counterbalance" their high term counts.

Generally, with respect to the distribution of words in a corpus, i.e., across documents, the most influential characterization comes from a model called **Zipf's law**, which states that each term's frequency is inversely related to its rank in the frequency table. In other words, the most common word appears twice as often as the second-most common, three times as often as the third-most common, and so on – which implies that term frequencies within corpora exhibit a power law (Manning et al. 2008, p. 89).

Finally, to conclude this subsection, it is worth to reference the interested reader to Zobel and Moffat (1998), who present other common weighting schemes (for instance, the Okapi BM25 function, which is very popular amongst practitioners).

3.3.2 Weight Aggregation, Sentiment Scores, and Volatility-Impact-Based Term Weighting

Having obtained weighted term frequencies, the next step is to aggregate the separate term counts/weights into a single document-wide score for each 10-K* filing. Following the approach described in Manning et al. (2008) and adopted in a slightly modified fashion in

Jegadeesh and Wu (2013), the k-th LM-category score for document i is calculated as a simple sum of the (weighted) term frequencies. For instance, using `TFIDF` results in the following score construction:

$$S_i^k(\texttt{TFIDF}) = \frac{1}{L_{i,k}} \sum_{j=1}^{J_k} tfidf_{i,j}^k, \tag{3.8}$$

where the first term scales the score by document length (denoted by $L_{i,k}$ and measured as element-wise sum of the i-th row in the k-th DTM). Replacing $tfidf_{i,j}^k$ with relative, log-transformed, or maximum-scaled counts, one can similarly obtain $S_i^k(\texttt{RF\_IDF})$, $S_i^k(\texttt{WF\_1PLOG})$, $S_i^k(\texttt{WF\_LOG1P})$, $S_i^k(\texttt{WFIDF\_1PLOG})$, $S_i^k(\texttt{WFIDF\_LOG1P})$, and $S_i^k(\texttt{TFMAX})$, respectively.

However, as Jegadeesh and Wu (2013, p. 7) point out, "although idf weights have an appeal in other contexts, there is no particular reason that the frequency of occurrence of a word in documents should be related to market's perception of its impact". Therefore, they propose to estimate the score using a different, market-based adjustment of term frequencies. To illustrate this idea, it is helpful to recall the definition of `TFIDF`: $tfidf_{i,j}^k = tf_{i,j}^k idf_{j,k}$. This can be read as a simple two-component product of the term frequency and the IDF-weight for the corresponding term j. The idea of Jegadeesh and Wu (2013) basically seeks to replace the second part of this product: instead of using a corpus-based weight, one could alternatively find a market-based weight, that might prove more suitable in modelling market reactions to textual contents. As Jegadeesh and Wu (2013) modelled CAR subsequent to 10-K submissions, "market-based" weighting in their setting seeks to assign a higher weight to those words in the vocabulary, whose appearance in past 10-K documents led to higher post-filing CAR. Jegadeesh and Wu (2013) call these *word power weights*. In a similar framework, in this thesis I will apply such a market-based term weighting scheme on the basis of the impact that words from sentiment "category" k had on realized volatility "1-week" after submission[40].

The next step is to estimate these market-based weights (henceforth labelled by w_j) in a linear regression framework based on a large number of in-sample, "training" 10-K* filings from the past. As such, it is useful to firstly define a sentiment score calculated on the basis of volatility-impact-based term weighting (`VIBTW`) in a similar fashion as it was achieved in section 3.3.1:

$$S_i^k(\texttt{VIBTW}) = \frac{1}{L_{i,k}} \sum_{j=1}^{J_k} tf_{i,j}^k w_j \tag{3.9}$$

Under the assumption that `VIBTW` scores linearly relate to post-filing realized volatility (denoted by `PFRV` or $\sigma_{t,i}^{PF}$), one can write and expand as follows:

[40] The definition of post-filing realized volatility, denoted by $\sigma_{t,i}^{PF}$ and henceforth abbreviated by `PFRV`, is described in detail in the subsequent section 3.4.

$$\begin{aligned}\sigma_{t,i}^{PF} &= a + bS_i^k(\texttt{VIBTW}) + u_{t,i} \\ &= a + b\left(\frac{1}{L_{i,k}}\sum_{j=1}^{J_k} tf_{i,j}^k w_j\right) + u_{t,i} \\ &= a + \left(\sum_{j=1}^{J_k} rf_{i,j}^k bw_j\right) + u_{t,i} \\ &= a + \left(\sum_{j=1}^{J_k} rf_{i,j}^k B_j\right) + u_{t,i} \end{aligned} \tag{3.10}$$

where the notation indicates that document i was filed at time t. The crucial element in (3.10) is coefficient $B_j \equiv b \cdot w_j$. If this equation is estimated with OLS[41], one obtains estimates $\hat{B}_j$ instead of the weight estimates $\hat{w}_j$ which are seeked for, as the latter are scaled by the constant b. In order to attenuate potential distributional distortions arising from this, and following Jegadeesh and Wu (2013), I will standardize ("z-transform") the estimated coefficients so as to obtain estimated word weights:

$$\hat{w}_j = \frac{\hat{B}_j - \bar{B}_j}{\text{Std.Dev.}(\hat{B}_j)}, \tag{3.11}$$

with $\bar{B}_j$ being the mean estimate across all J_k words and the denominator denoting the standard deviation of the $\hat{B}_j$'s around that mean.

As indicated, the estimated $\hat{w}_j$ in equation (3.11) will be based on a subset of the available data, which is used to "learn" the volatility-based weighting scheme (i.e., a *training* set). The obtained estimates will then be used in combination with out-of-sample term counts (which come from "upcoming" 10-K*s) so as to get scores $S_i^k(\texttt{VIBTW})$ for the 10-K* filings in this so-called "test" set. Mathematically, index i in equation (3.10) will **not** range from 1 to N but rather to $N_1 < N$, where N_1 indicates the size of the training set. Similarly, the learned weights will be applied to out-of-sample 10-K* documents with $i = N_1 + 1, N_1 + 2, \ldots, N$. For the ease of notation, I will denote the size of the test set as $N_2 = N - N_1$.

In this context it is necessary to highlight that, due to the fact that estimated $\hat{w}_j$ instead of "true" weights w_j are used, the score of newly examined filings, $S_i^k(\texttt{VIBTW})$, will be measured with error. However, as Jegadeesh and Wu (2013) describe, because of the high number of regressors embedded in the estimation of equation (3.10) and (3.11), the calculation of scores is not likely to be highly affected by this measurement error. In

[41] Note that the summation index encompasses all words j that are element of the respective LM dictionary k. For instance, for the negative word lexicon, this implies that equation (3.10) will be a model with $J^N = 882$ regressors.

particular, individual miscalculations per word are likely to offset each other on an aggregated score-level, especially when estimation is performed on long(er) sample periods (i.e., large N_1)[42].

3.4 Measuring Post-Filing Volatility

Using the impact on post-filing volatility to train the term weights in equation (3.10), one needs to define how this variable is measured. As volatility per se is latent, it is not an observable quantity; thus, one needs a suitable ex-post proxy variable (Brownlees et al. 2011). The most common choices are realized volatility measures (for *daily* data this imposes the additional requirement of availability of intra-day data but it is more tractable on a weekly time frame), squared returns (which have to be shown to be rather noisy, see Patton and Sheppard (2009)), or the (intra-day) range of log prices (loosely spoken, the difference between highest and lowest price).

In this thesis, I will opt for a measure of weekly realized volatility in an event-study-like fashion, where the event window starts with the submission of the filing and ends within a single business week (i.e., four days) after the publication[43]. In other words `PFRV` is measured as the (natural logarithm of) sample standard deviation of daily returns within the first τ days after the publication of the 10-K*. Analytically one can formulate as follows:

$$\sigma_{t,i}^{PF} = \ln\left(\sqrt{\frac{1}{|\tau|-1}\sum_{s=t}^{\tau}\left(R_{s,g}-\overline{R}_{[t,\tau],g}\right)^2}\right), \tag{3.12}$$

where the notation suggests that report i was filed by submitter g at date t, $|\tau|$ is the size of the event window in days (here: $\tau = t+4$ implying $|\tau| = 5$[44]), $R_{s,g}$ is the log-return of company g's stock closing price on day s (i.e., $R_{s,g} = \ln(P_s) - \ln(P_{s-1})$) and

[42] In fact, Jegadeesh and Wu (2013, pp. 14, 15, 39) examine the "stability" of scores using two/three non-overlapping sub-periods. Indeed, correlation between the scores is significant, implying they are "stable" over time. This indicates that variation in scores is likely to come from variation in the "true" scores rather than from measurement error.

[43] Note that squared returns and absolute returns will be considered as alternative volatility proxies in robustness checks (section 7.2).

[44] The choice of the estimation window is motivated by two reasons: firstly, measuring daily volatility requires – unless intra-day or tick data is available – the use of a much noisier proxy of true volatility, such as squared returns. In this context Engle and Patton (2001, p. 240) highlight that longer sampling frequencies can (partly) circumvent this problem by using more reliable realized volatility measures. The second argument in favour of a slightly longer post-filing window is the fact that 10-K* reports are usually quite comprehensive and long. As it will be indicated in the data description (section 5) the median 10-K* has 49,117 words, implying that even institutional investors need some time to process the information contained in the filings properly. Loughran and McDonald (2011) experimented in shrinking their event window for absolute excess returns and experienced the results becoming insignificant when the event window was narrowed to one or two days post-filing, respectively; leading them to conclude that "The document length would require the average investor some period of time to absorb the information" (Loughran and McDonald 2011, p. 53). However, the choice of the post-filing window needs to be a balanced one; a too long post-filing window leaves larger room for other important events

$\overline{R}_{[t,\tau],g} = 1/|\tau| \sum_{s=t}^{\tau} R_{s,g}$ is the average log return of the stock price of company g in the time window $[t, \tau]$.

For the sake of completeness, it is worth mentioning that more sophisticated volatility proxies exist. For instance, Raviv (2012) presents three alternative volatility measures (usually thought to take use of intra-day data), that can also be employed for a weekly volatility measure if suitable data is available. They are "range-based" measures, that mainly use open-high-low-close (OHLC) data so as to capture the dispersion of price movement instead of solely relying on close-to-close (log-) returns[45].

to occur, which could have substantial volatility impact and hence bias the relative importance of the (textual part of the) 10-K*.

[45] Interested readers are referred to Patton and Sheppard (2009) for a profound analysis of how imperfect volatility proxies affect the forecasting task and subsequent statistical and accuracy testing procedures.

4 Research Framework: Cross-Sectional Volatility Model

In this section I will present the second pillar of the research design of this work. It deals with the usage of the sentiment scores calculated in the previous section 3 for out-of-sample volatility prediction purposes, embedded in a set of seven research hypotheses. Moreover, one key model input is introduced: resembling the idea of volatility clustering and persistence, pre-filing realized volatility will be used as a predictive variable. In addition, control variables and their construction are presented. Finally, the model specification is explained in further detail.

4.1 Pre-Filing Realized Volatility as Predictive Variable

Before relating `PFRV` to the qualitative variables extracted from 10-K* text, it is important to highlight the most "influential" right-hand side variable that will be used in the models. It is known from empirical literature that return volatility (or more generally, squared or absolute returns) is highly persistent, especially if sampled from longer frequencies. Following Kogan et al. (2009), Tsai and Wang (2016), and Rekabsaz et al. (2017), who all used past volatility as a baseline predictor for post-10-K-filing volatility, I will use pre-filing realized volatility as a predictor in all models. Kogan et al. (2009, p. 275) in that context highlight that "Volatility is, generally speaking, not constant, yet prior volatility [...] is a very good predictor of future volatility". This resembles a very basic "auto-regressive" model in realized volatility; obviously, however, with an abuse of terminology, as the analysis is performed on the cross-section of 10-K* filings. Pre-filing realized volatility will be measured in the same manner as its ex-post counterpart (i.e., using the natural logarithm of the in-sample standard deviation as in equation (3.12)), with the modification of $s = t - 5$ and $\tau = t - 1$ so as to capture volatility in the week *before* the filing. In tabulations and the result analysis further on, pre-filing realized volatility will accordingly abbreviated with the variable name `PreFRV`.

With respect to the motivation of including a quantitative "time-series"-like forecast into a predictive model, Rekabsaz et al. (2017) show that volatility models which incorporate both quantitative and qualitative information from 10-K* reports carry the most explanatory power (compared to benchmark models that use either of the two sources separately). Indeed, the main research interest of this thesis is not exclusively focussed on the question whether text embedded in corporate filings provides value for volatility prediction purposes; rather, I seek to investigate whether textual sentiment provides value *added*, i.e., offers incremental value when used in addition to a historical quantitative predictor.

4.2 Hypotheses Development: Connecting 10-K* Text to Realized Volatility

As indicated in section 3, the main approach of incorporating 10-K* sentiment into volatility prediction is to learn term weights on the basis of past volatility impact from a training set of 10-K*s, and apply those weights to term frequencies of words contained in a LM-dictionary that is intended to measure sentiment in a particular category. The combination of term frequencies of words from sentiment "class" k in an out-of-sample filing with their corresponding class-k weights learned from the training sample N_1 will lead to the creation of a k-related sentiment measure, which is designed to extract a specific aspect in the management's writing style and potentially be connected to `PFRV`. Therefore, with respect to the textual analysis, I present the following seven hypotheses about how textual information embedded in annual reports is related to post-filing volatility.

H1: Negative sentiment (`NEG_SENT`) embedded in 10-K*s, reflected by higher S_i^N(`VIBTW`), is associated with higher `PFRV`.

This hypothesis relates to Kothari et al. (2009), who were among the first to test textual impact on volatility. In a similar manner, I hypothesize that negative tone in corporate filings will lead market participants to perceive the filing company as more risky, implying, ceteris paribus, an increase in the post-filing volatility of stock returns.

H2: Positive sentiment (`POS_SENT`) embedded in 10-K*s, and measured by S_i^P(`VIBTW`), is associated with `PFRV`.

Following the findings of Kothari et al. (2009), who find that positive tonality in news disclosure by management, analysts, and business press decrease quarterly volatility after the release of such information, I hypothesize that a similar relationship holds for weekly post-filing volatility and sentiment contained in 10-K* filings. However, as indicated in section 3.2, the interpretation of coefficients assigned to `POS_SENT` requires increased awareness and carefulness: as potential negation of positive phrases is neglected, `POS_SENT` in fact might be negative sentiment which is formulated using negated positive terminology[46]. This would imply that `PFRV` would, ceteris paribus, increase for filings with larger degree of positive sentiment, thereby potentially resembling results in Antweiler and Frank (2004), who evidenced bullishness in message boards to lead to larger volatility after the message was posted. Therefore, no prediction will be made on the sign of the relation between positive sentiment and post-filing volatility.

[46] In their meta study on textual analysis in finance, Loughran and McDonald (2016) even suggest that "Unless a study can convincingly resolve the problems of negation, positive sentiment is best left untested" (Loughran and McDonald 2016, p. 1217) and, therefore included them in their 2011 study "more in the interest of symmetry" (Loughran and McDonald 2011, p. 45).

H3: Assertiveness (ASSERT) in the management's writing style, reflected by higher S_i^{SM}(VIBTW), is associated with lower PFRV.

This hypothesis relates to A. H. Huang et al. (2014), who tested how the assertiveness in the textual part of analyst reports affected abnormal returns in the equity markets. As they document with reference to psychology literature, investors perceive assertive and confident communicators as "more accurate, competent, and credible" and thus assume that they "convey information signals with greater precision" (A. H. Huang et al. 2014, p. 2157). Therefore, I hypothesize that market participants will perceive the filing company as less risky, if the management applies a higher degree of conviction and assertiveness in writing the annual report.

Regarding the measurement of assertive tone, I will follow A. H. Huang et al. (2014) and use the strong modal sentiment list from LM as an indicator for assertive and persuasive language.

H4: Usage of language related to uncertainty (UNCERT) in the 10-K* filing, reflected by higher S_i^{U}(VIBTW), is associated with higher PFRV.

In contrast to H3, I hypothesize that if tone in 10-K* documents indicates uncertainty, investors will be more alert and interpret the information conveyed in the filing more cautiously. This increased incertitude about the future performance of the company will affect their valuation, and thereby increase the risk perceived – leading to larger PFRV. When creating their dictionaries, Loughran and McDonald (2011) tested this relationship on data starting five days after the filing and covering the whole subsequent year. Indeed, they found a positive association between volatility and frequency of words from the uncertainty list. In this context, it remains to be tested whether this association is also evident in a shorter post-filing window as well as under usage of VIBTW-weighting schemes.

H5: Litigious language (LITI) embedded in corporate filings, reflected by higher S_i^{L}(VIBTW), is associated with higher PFRV.

Under the assumption that the management of the filing company uses the 10-K* filing to be precautionary and inform (or forewarn) investors about potential future lawsuits or required legal action, this increased uncertainty about future cash-flows related to such juridical incidents will negatively influence stock valuation and perception about risk/volatility (i.e., increase PFRV). As for the uncertainty word list, Loughran and McDonald (2011) evidence this relationship for the post-filing year, excluding the first five days after the filing, leaving room to test the robustness of their findings in a shorter post-filing window and in combination with an alternative term-weighting scheme.

H6: Longer, and consequently less readable annual reports (measured by the natural logarithm of gross file size, GFS), are associated with higher PFRV.

This hypothesis is motivated by deliberations in Loughran and McDonald (2014, p. 1644) who state that "better written documents produce less ambiguity in valuation, as reflected by the lower price volatility of the stock in the period immediately following the filing." Thereby, more readable filings are associated with greater certainty and hence can be viewed as logical opposite to H4 (uncertainty).

A. H. Huang et al. (2014, p. 2157) in their study using post-filing CAR, from a signal-versus-noise perspective, additionally highlight that "Longer annual reports are less readable and harder to process. Therefore, it is reasonable to assume that a more concise report is easier to process and likely to receive more attention than a longer report, resulting in a greater price reaction." As the authors focus on returns rather than return volatility, it is unclear whether this idea can be extended to the latter. One could assume that a similar relation might hold at least for trading volume, which in turn is often positively related to volatility, making the channel of cause and effect an indirect one. By this line of thought, short (and thus readable) 10-K*s would lead to enhanced trading activity and thereby *higher* post-filing volatility. Vice versa, longer and less readable 10-K*s would imply lower trading activity and lower `PFRV`.

Supporting this conjecture, and relating to the obfuscation theory provided in F. Li (2008, p. 221) and briefly described in section 2.1.1, one could also hypothesize that management tends to produce longer reports so as to dilute and hide bad news in the filing. If this attempt succeeds (i.e., the filing is indeed longer and less readable) the usage of inflated language could disguise negative news, thereby inducing lower `PFRV`. This explanation, however, lacks to cover cases in which report size does not coincide with the presence of bad news (e.g., a report might be longer simply due to the complexity of the business the firm operates in, as indicated in Loughran and McDonald (2016)); in this context, there would be no incentive to disguise information by verbose language, thereby breaking the significance of the obfuscation theory. Therefore, I hypothesize that the first effect dominates and more (less) readable 10-K*s will induce lower (higher) `PFRV`.

In terms of measurement of length and readability, I follow Loughran and McDonald (2014), who suggest the (logarithm of) gross file size of the filing document as a potent proxy for readability, which at the same time allows to easily circumvent the issues arising from measures that are influenced by the parsing procedure applied by the researcher.

H7: The management's focus on financial topics (denoted by `FIN`) is associated with lower `PFRV`.

This hypothesis is motivated by Amel-Zadeh and Faasse (2016, pp. 1-2) who state that "if information is abstract and heavily loaded with statistical and quantitative data, people tend to underweight it in their decision making". Following this notion, embedding qualitative information in a glut of financial figures, ratios, and percentages will engender that market

participants perceive this content as less informative, thereby reducing the relevance of the textual part and leading to lower PFRV.

Similarly to A. H. Huang et al. (2014), I will measure FIN as a function of the number of appearances of financial symbols. In fact, in the cleaning process of the corpus, I convert symbols to their corresponding word equivalents (e.g., % becomes percent, $ becomes dollar, and so on). I then construct the FIN variable as the log of one plus the sum of the term frequencies for the following words: percent, dollar, euro, yen, pound, and franc (thereby covering occurrences of most commonly used currencies that could potentially arise within a filing).

4.3 Control Variables

In order to control for other potential factors that influence PFRV, I consider some additional independent variables that have been shown to be successful predictors of volatility:

(1) SIZE: natural logarithm of total assets at the fiscal year-end preceding the 10-K* filing year. Smaller firms are considered more risky due to the fact that their assets and projects are relatively undiversified (Kothari et al. 2009).

(2) BTM: natural logarithm of the book-to-market ratio at the fiscal year-end preceding the 10-K* filing year. Firms with higher book-to-market ratios are on average expected to be more risky, as market participants assign lower market valuation (denominator) to companies if they perceive their future cash flows to be rather uncertain (Kothari et al. 2009).

(3) TRVOL: natural logarithm of the median stock trading volume (number of shares that have been traded) in the business week (i.e., five days) preceding the 10-K* filing date. This is motivated by model specifications in Antweiler and Frank (2004) as well as Tetlock (2007), who on the basis of findings in Jones et al. (1994) include trading volume as predictor for stock volatility. The latter evidence that the positive relationship between volume and volatility is driven by the sheer frequency of transactions (i.e., the number of shares traded generates volatility, regardless of the *size* of the transaction).

(4) VIX: median level of the CBOE VIX in the business week (i.e., five days) preceding the 10-K* filing date. This variable is included as a measure for market-wide volatility and serves to better isolate the "idiosyncratic" part of firm-level volatility.

(5) LEVER: financial leverage (i.e., total liabilities divided by total assets) at the fiscal year-end preceding the 10-K* filing year. More levered companies are perceived as more risky by market participants, therefore being related to higher stock return volatility (Kothari et al. 2009).

(6) `YRDUMMY`: 18 dummy variables indicating the year in which the 10-K* was filed. This variable is included to control for annual trends.

(7) `MTHDUMMY`: eleven dummy variables indicating the month in which the 10-K* was filed. This variable is included to control for seasonality and/or other market anomalies within a calendar year (e.g., turn-of-the-year / "January" effect).

(8) `WEEKDAYDUMMY`: four dummy variables indicating the weekday on which the 10-K* was filed. This variable is included to control for weekday effects as well as to account for potential disruptions of business weeks due to weekends.

(9) `MONTHDAYDUMMY`: 30 dummy variables indicating the day of the month on which the 10-K* was filed. This variable is included to control for "turn of the month" effect.

(10) `SECTORDUMMY`: 65 dummy variables indicating the first two digits of the SIC code of the filing company. This variable is included as a measure for industry differences.

(11) `10KDUMMY`: 12 dummy variables indicating the 10-K filing type. This variable is included so as to account for potential different volatility responses due to the nature and type of the filing; for instance, the explanatory power of text in an amended 10-K might be different from a "regular" 10-K.

4.4 Bringing Everything Together: Model Specification

Using the definition of `PFRV` from section 3.4 (equation (3.12)) and combining it with the sentiment scores calculated using equation (3.10) presented in section 3.3 as well as the control variables (section 4.3), I aim to test the hypotheses expressed in section 4.2 with a linear model specification, which will be explained in this subsection.

For illustrative purposes, Figure 2 additionally describes the research design in graphical manner. As is indicated in the legend, the orange area with diagonal pattern represents three years of rolling training data which is used to estimate the regression in equation (3.10). The green dots represent the filing dates; the green area with horizontal pattern represents the post-filing time horizon - it is set equal to $[t, t+4]$ days for filings submitted in t so as to measure weekly volatility. The blue area with vertical pattern represents a varying time window that is used to construct independent variables, where applicable (for instance, for `PreFRV`, `TRVOL` and `VIX` it will be five days).

4.4.1 Linear Regression Model

The multivariate model attempts to describe post-filing realized volatility as a linear combination of different explanatory variables in a classical regression framework. As all

independent variables are known at the respective filing date, this model can actually be interpreted as a forecasting exercise. Moreover, as PFRV was modelled using a fixed post-filing time window, the linear model described below applies to the *cross section* of out-of-sample annual filings.

Analytically, the linear model can be described as follows:

$$\begin{aligned}\texttt{PFRV} =& \beta_0 + \beta_1\texttt{PreFRV} + \beta_2\texttt{NEG\_SENT} + \beta_3\texttt{POS\_SENT} + \beta_4\texttt{ASSERT} + \beta_5\texttt{UNCERT} + \beta_6\texttt{LITI} + \\ & \beta_7\texttt{GFS} + \beta_8\texttt{FIN} + \beta_9\texttt{SIZE} + \beta_{10}\texttt{BTM} + \beta_{11}\texttt{TRVOL} + \beta_{12}\texttt{VIX} + \beta_{13}\texttt{LEVER} + \\ & \sum_{l=14}^{31}\beta_l\texttt{YRDUMMY}_l + \sum_{l=32}^{42}\beta_l\texttt{MTHDUMMY}_l + \sum_{l=43}^{46}\beta_l\texttt{WEEKDAYDUMMY}_l + \\ & \sum_{l=47}^{76}\beta_l\texttt{MONTHDAYDUMMY}_l + \sum_{l=77}^{141}\beta_l\texttt{SECTORDUMMY}_l + \sum_{l=142}^{153}\beta_l\texttt{10KDUMMY}_l + e \end{aligned} \tag{4.1}$$

For the sake of readability, document (i) and date (t) indices are suppressed. Equation (4.1) will be estimated via OLS for all out-of-sample 10-K*s. The sample split into training and test set will cover three variants: a static, rolling, and extending training window. Details regarding the size of training and test set will be outlined in the next section, 4.4.2. Moreover, applying common hat notation to denote sample estimates for the coefficients in equation (4.1), with respect to the hypotheses presented above, I expect $\hat{\beta}_2$, $\hat{\beta}_5$, $\hat{\beta}_6$, and $\hat{\beta}_7$ to be positive and $\hat{\beta}_4$, and $\hat{\beta}_8$ to be negative. $\hat{\beta}_3$ is expected to be different from zero, with no prediction on the sign.

Referring to time-series practice, regressing "actual" values of volatility (proxied by realized volatility) against forecasted values of volatility is a common method to evaluate forecasting accuracy and is referred to as Mincer-Zarnowitz-Regression (henceforth MZ-regression). In the framework of this thesis, the univariate MZ-regression reads $\texttt{PFRV} = \beta_0 + \beta_1\texttt{PreFRV} + e$. The "suitability" of the basic "auto-regressive" model in volatility would then be evaluated by a joint test of H0: $\beta_0 = 0 \wedge \beta_1 = 1$ versus H1: $\beta_0 \neq 0 \vee \beta_1 \neq 1$ (Diebold 2017, p. 337). Of course, the MZ testing procedure can be applied for any sort of forecast and is not restricted to the basic case of using pre-filing volatility to explain post-filing volatility; in fact, in more sophisticated predictors from the GARCH-model family will be considered in robustness checks to the specification just presented (cf. section 7.3).

Extending the concept of the MZ-regression, equation (4.1) can be seen as an "augmented MZ-regression", whose main function is to provide evidence on whether textual variables extracted from corporate 10-K* filings can explain post-filing realized volatility for a given firm in the cross section, after controlling for other common factors that have shown to influence stock return volatility. In terms of forecast evaluation, a **fully** capable quantitative model would imply that all beta coefficients in equation (4.1), with the exception of β_1, are zero (Patton and Sheppard 2009; Violante and Laurent 2012), whereas in this work I will

be moreover interested in the significance of the other non-zero coefficients, and especially coefficients of those variables that relate to the textual part of the 10-K* filing.

4.4.2 Choice of Training and Test Set

Summarizing briefly, the research framework of this thesis bases upon `VIBTW`, which in turn grounds on three basic steps:

1. Estimating `VIBTW`-weights $\hat{w}_j$ for each word j in LM lexicon k using equations (3.10) and (3.11). This linear regression is estimated by OLS using the *training* sample, $N1$.
2. Applying the learned term weights $\hat{w}_j$ to term frequencies that are **outside** of the training sample, which involves multiplying them with relative term frequencies from 10-K*s in the testing sample (N_2). After this transformation, `VIBTW`-weighted term frequencies are aggregated into a sentiment score for category k (using equation (3.9)).
3. Using sentiment scores from out-of-sample 10-K*s to fit an augmented MZ-regression (equation (4.1)), which attempts to explain `PFRV`. This regression is estimated using the test set, N_2.

Thus, one needs to split the sample in such manner that a fraction N_1 is used for weight estimation and the remaining fraction (N_2) is used for score calculation and MZ-regressions[47]. Regarding the choice of N_1 and N_2, three alternative specifications will be tested in this thesis:

1. Static Window: the training window will be the same for each of the five out-of-sample years ranging from 2013-2017, implying that N_1 will encompass all 10-K*s from 1999 up to (including) 2012[48].
2. Rolling Window: the natural extension to a static 14-year in-sample period is to make that subset roll along for each year in the test set. For instance, `VIBTW`-sentiment scores in equation (4.1) for out-of-sample year 2013 will be based on weights derived from all filings from 1999 to 2012; whereas for 2014 they will be derived from 2000-2013; for 2015 they will be derived from 2001-2014; and so on.

[47] Note that the sample split is not exclusive to `VIBTW`; it also required for calculating inverse-document-frequencies.

[48] The choice to set the training window to 14 years, i.e., about 75 percent of the whole sample, is motivated by two reasons: firstly, although the choice is still arbitrary, many applications adopt a full sample split of 70 versus 30 or 80 versus 20 percent; secondly, the 14-year subset is large enough to avoid singularity and collinearity problems in estimating equation (3.10) and yet leave a large enough test sample to be able to fit the regressor-rich model in equation (4.1).

3. Extending Window: another alternative is to fix the starting point of the training set but let the ending point roll along, thereby making the in-sample window grow. As a consequence, for each year in the test set, all available information prior to the filing year is used to estimate `VIBTW`-weights. Thus, for 2013 the training window will (equivalently to methods 1 and 2 above) be 1999-2012, while for 2014 it will be 1999-2013, for 2015 it will be 1999-2014, etc.

5 Data and Sample Description

This section will initially outline the data collection and sample creation process. Then it will provide insights regarding the composition of the final corpus containing 46,483 annual filings as well as descriptive statistics about those filings. Finally, general textual features about the 10-K* corpus as well as descriptive statistics about the variables used, will be depicted.

5.1 Data Collection, Sample Formation and Matching

The aim of this section is to outline the data collection and sampling procedure. Firstly, all 10-K* filings filed with the SEC between 1994 and 2017 were downloaded from `https://sraf.nd.edu/`, a software repository maintained by Loughran and McDonald at the University of Notre Dame, which was developed during their works on 10-K* filings (Loughran and McDonald 2011; Loughran and McDonald 2014). They provide a collection of what they refer to as "stage one parse files", i.e., filings that were already parsed and cleaned from mark-up language tags such as HTML or XBRL as well as "non-textual" information embedded in the filings (e.g., encoded graphics, tables, files, pdf-documents, etc.)[49]. Thereby, Loughran and McDonald make available a much more structured corpus of 10-K* filings, as they ease the researcher's burden of dealing with "raw" downloads from the SEC EDGAR web-page.

The initial, full sample contains 295,746 10-K* filings, submitted by 40,264 firms in the time period between 01/01/1994 and 12/31/2017. However, due to the necessity of collecting time series of prices for volatility calculation as well as other control variables (as pointed out in the description of the research design in sections 3 and 4), the sample of usable 10-K*s reduces in the data matching procedure. Table 1 describes the sample construction process in greater detail. As one can see, for roughly sixty percent of observations there are stock tickers available, based on which the price series can be downloaded[50]. The time series of stock prices and trading volumes were obtained from the Yahoo! Finance database in the time period between 06/05/2018 and 06/10/2018 and was available for roughly 42 percent of the identified tickers. In addition, the sample diminishes further when matching the 10K* corpus with data for the corresponding control variables, which were collected from

[49] All parsing details are provided at `https://sraf.nd.edu/data/stage-one-10-x-parse-data/`. One important note regarding tables in the filings: some companies (for formatting purposes) apply tables to demarcate separate sections of the filing. Thus, by simply stripping out all text that occurs between mark-up tags `<TABLE>` and `</TABLE>`, one might actually lose relevant paragraphs containing textual contents. Thus, Loughran and McDonald only eliminate those tables where the ratio of numeric characters to total (i.e., numeric plus alphabetic) characters is larger than ten percent.

[50] The matching of CIK (*central index key*), the identifier used by the SEC, and stock tickers is based on the following list as of 06/11/2018: `http://rankandfiled.com/#/data/tickers`.

the Compustat database, available via WRDS[51]. The final sample of 10-K*s after matching the filings with Yahoo! Finance prices and Compustat WRDS data contains 46,483 filings from 3,736 companies for the time period from 01/06/1999 till 12/29/2017.

5.2 Sample Composition: Filing Types and Timing

Some distributional properties of the 46,483 available filings are displayed in Figure 3, where the bars represent the number of filings in the respective year, month, day of the month, or weekday; the coloured stacks *within* bars indicate the filing type[52]. As one can observe in the upper-left panel (A), the number of filings follows an upward trend over the inspected time period, with the count of 10-K* submissions increasing by about 47 percent (comparing 2017 with 1999). On average, there are 2,446 filings available for each year in the sample (median: 2,515). Further, the upper-right panel (B) indicates that the vast majority (namely more than 41 percent) of 10-K*s are submitted in March. This is mainly due to the combination of most companies having a fiscal year deadline that coincides with the calendar year and the requirement of the SEC to submit the filing within 60-90 days after fiscal year end[53]. The lower-left graph (panel (C)) displays the distribution over days of the month; the two humps give rise to the conjecture that firms with fiscal years ending on either the 15th or the 30th (31st) day of the month will tend to submit filings just a few days before (or on) the deadline day. The lower-right panel (D) in Figure 3 shows the distribution of filings across weekdays, with Fridays leading ahead as submission weekday[54]. These observations provide support for the choice to include year and month dummy variables to control for trends and seasonality as well as weekday- and day of month indicators to account for potential "weekend" or "turn of month" effects (or, phrased more generally, to control for so-called "calendar effects" in the filing behaviour).

5.3 Descriptive Statistics About the Corpus

This subsection provides descriptive statistics about the textual aspects of the 10-K* corpus used in this thesis. The corpus composition over years is displayed in Table 3; in addition, the total, mean, and median number for both tokens and types (i.e., unique tokens) is presented. As can be observed, both average and median word count increased in the past

[51] `https://wrds-web.wharton.upenn.edu/wrds/`, accessed on 06/13/2018.

[52] The stacked bars are hardly distinguishable due to the fact that the large majority of reports are of "standard" type 10-K, as indicated by the darkest-blue stacks, which dominate across years, months as well as days of month and week. In fact, 42,762 of the 46,483 reports (i.e., 91.99 percent) are either "vanilla" 10-K or amendments thereof (10-K-A). The detailed decomposition of filing types is provided in Table 2, and potential differences in post-filing volatility that are attributable to different filing types will be accounted for by adding filing type dummy variables in the regression tests.

[53] See `https://www.sec.gov/fast-answers/answers-form10khtm.html` for details regarding filing deadlines.

[54] Note: No 10-K* were filed on Saturdays or Sundays.

19 years – as was already evidenced in Kogan et al. (2009) and Rekabsaz et al. (2017). The former especially investigated upon the jump in word counts occurring in 2002-03: this increase in verboseness can potentially be attributed to the enactment of the Sarbanes-Oxley Act in 2002.

Moreover, Table 4, presents the top-5 words for each LM-lexicon over the whole sample period (1999-2017) as well as three sub-groups. It appears that this tail of the distribution of word counts is stable over time, as most of the five top-words appear in each of the three sub-groups (very often even in the same ranking). This gives rise to optimism when applying weights that are based on past term frequencies to out-of-sample count observations.

5.4 Descriptive Statistics About the Variables

Concluding this section, some descriptive statistics about the variables of the linear model described in equation (4.1) will be displayed and commented. Table 5 provides standard distributional figures such as the number of observations, the sample mean, sample standard deviation, 25/50/75 percentiles, as well as minimum and maximum value observed in the sample.

One can see that `PFRV` (expressed in logarithmic form) is ranging from .015 percent to 1,091 percent. As the latter clearly appears to be an outlier, the average realized volatility in the full sample is in range on a weekly level and amounts to 2.19 percent, whereas the median is slightly lower (2.11 percent). For the pre-filing counterpart, the figures are highly similar, yet a little lower, thereby indicating that `PreFRV` **might** – on a univariate basis – be a too optimistic predictor for `PFRV` by producing too small post-filing volatility estimates. Further details on the distribution of `PFRV` and `PreFRV` as well as their univariate relationship will be provided in the section on univariate results (cf. section 6.1.1).

With respect to the subsequent rows of Table 5, which are related to the textual content of the 10-K*, one needs to highlight that variables `NEG_SENT`, `POS_SENT`, `ASSERT`, `UNCERT`, and `LITI` are `VIBTW` scores based on estimated weights that stem from the **full** sample (no in- versus out-of-sample split was applied in computing the weights and scores). Note that sentiment *scores* are displayed rather than weights, as the latter, due to the z-standardization described in equation (3.11), have zero sample mean and unit variance (standard deviation) by construction. As sentiment scores per se are difficult to interpret in terms of their scale, Figure 4 exhibits six histograms for the five sentiment scores as well as the readability measure for illustrative purposes and ease of interpretation. The distribution of scores appears close to Gaussian and is certainly symmetric, while readability appears to be slightly more skewed (to the left) as well as mesokurtic (compared to the normal distribution). The last text-related variable captures the focus on financial topics (`FIN`); considering that it is calculated as the logarithm of one plus the

term count of financial keywords, one can see that the term count ranges from zero up to 3,363, with the mean frequency being equal to 493 financial expressions per filing.

The remaining variables serve as control variables to account for a potential confounding effect; for instance, smaller and higher levered firms have been shown to be more risky. In terms of size (measured as logarithm of the firm's total assets), the sample ranges from small firms with an asset base of three million USD on the lower end (ignoring one percent of outliers) to large financial institutions like Citigroup or JPMorgan Chase with assets of multiple trillion USD. The latter represents the maximum in the sample with the 10-K* filed in 2017 and total assets of 2.49 trillion USD. The average (median) company in the sample operates with assets of 670 (600) million USD. With respect to book-to-market-ratio (`BTM`), the distribution appears largely driven by outliers (minimum of zero and maximum of 4,040 for a mean (median) of .75 (.49)). So as to dampen this distributional distortions, this variable is used in logarithmic form, making it more Gaussian yet leaving visible leptokurtosis. A similar picture is obtained for variable `TRVOL`, measuring the median trading volume in the week preceding the 10-K* filing. The measure ranges for a single share to 511 million shares traded, with the average being 1.23 million shares traded a day. Applying the logarithmic transformation helps to standardize the variable, making it close to normal (skewness: -.4, kurtosis: 2.89). Control variable `VIX` serves to account for market-level volatility, and it should be highlighted that the descriptive statistics in Table 5 do not apply for the VIX during the whole sample period (1999-2017). Instead, the variable represents the median value of the VIX in the week preceding the filing date. As indicated, `VIX` on average had a value of 19.68 (median slightly lower with 17.86), reaching its minimum (9.43) during the very calm market period in July 2017 and its maximum (72.67) in late November 2008 as a consequence of the crash in the housing market and the subsequent start of the financial crisis. Finally, the filing company's leverage ratio is considered as control variable. Among the 46,483 corporate filings available for analysis, `LEVER` on average was 52 percent, and was within the economic and accounting "boundaries" of zero (full-equity) and unity (full-debt).

6 Results

This section will display empirical results for the uni- and multivariate analyses on the 10-K* corpus presented in the previous section 5. Additionally, a brief analysis of a basic MZ-regression of `PFRV` versus `PreFRV` will be shown. Finally, linear regression results in a multivariate setting for the two model specifications and three training split methodologies will be discussed.

6.1 Univariate Analysis

6.1.1 Pre- and Post-Filing Realized Volatility

As a starting point, I will consider an univariate analysis, which aims to test the accuracy of a simple time-series baseline: it shall be evaluated how well the level of volatility just *before* the filing is submitted is able to explain the value of volatility *post*-filing. Table 6 displays the correlation coefficient of `PFRV` with `PreFRV` as well as other volatility measures, which will be used in robustness checks (see the Table footnote on the variable descriptions and the corresponding section in this work). As expected, the correlation between pre- and post-filing realized volatility is positive and highly significant in the cross section of 46,483 annual reports analysed. Figure 5 extends the idea of correlation analysis and displays the scatterplot of `PFRV` versus `PrefRV` as well as the fitted MZ regression line. As it is observable, however, the goodness-of-fit is rather poor, although it seems mainly driven from outliers visible on the plot. A R-squared statistic of .34 confirms this observation and gives rise to call for an extended model with more control variables, which should be better able to capture cross-sectional variation of weekly realized volatility after the filing of a 10-K* report. In addition to the scattermatrix, Figure 5 also displays the marginal distribution of `PFRV` and `PreFRV` (see the blue-shaded histograms along the axes). Both variables follow a bell-shaped distribution in their logarithmic form. This observation with regards to the distribution of volatility in the cross section of firms was already evidenced in Kogan et al. (2009), Tsai and Wang (2012), Tsai and Wang (2013), and Tsai and Wang (2016). Therefore, logarithmic versions of the volatility measures will be used for the multivariate analysis as well. Additionally, Figure 6 shows the forecast error of the "auto-regressive" model in realized volatility (i.e., using pre-filing realized volatility to predict post-filing realized volatility). The delta series distribution is bell-shaped, almost symmetric (skewness: -.065) and slightly leptokurtic (kurtosis: 4.94). The inspection of the forecast error moreover reveals that on an aggregated delta between pre- and post-filing realized stock return volatility is negligibly small (median forecast error: .07 percentage point, mean forecast error: .2502 percentage points), yet indicates that `PreFRV` slightly overestimates `PFRV`.

6.1.2 Correlation Analysis

Continuing the univariate analysis, I seek to explore how both text-related and control variables correlate with post-10-K*-filing stock return volatility. Table 7 shows the correlation matrix for all variables used in this thesis; the lower triangular (including the main diagonal) of the matrix displays Pearson coefficients, while the upper triangular is filled with the corresponding p-values to test for significance.

Again, it needs to be pointed out that correlation coefficients and p-values are estimated from the whole sample. This constitutes a factor of special importance for the textual variables, which theoretically require a sample split into in- and out-of-sample fraction, but are computed on a full-sample basis so as to ensure comparability as well as a sufficiently large number of observations. The correlation coefficients of `NEG_SENT`, `UNCERT`, and `LITI` with `PFRV` are positive and significant and thus seem to confirm hypotheses 1, 4, and 5, respectively. However, the three variable correlate among each other as well, to some extent potentially reflecting the overlap in the three word lists. Similarly, hypothesis 7 seems confirmed on an univariate basis, namely that higher focus of financial topics coincides with cases of lower post-filing realized volatility. However, hypothesis 3 appears to be rejected by the correlation analysis, as reports in more assertive language seem to induce larger `PFRV`. A potential explanation might read as follows: assertiveness per se might be neither positive nor negative. Thus, the fact that the management applies language of conviction might simply be an *amplifier* that helps market participants to interpret the signals conveyed in the 10-K*, be it positive or negative ones. Hence, the "relevance" of news might, ceteris paribus, appear large and thus induce trading activity and potentially increase volatility (equally so for both positive and negative news). Moreover, relating to hypothesis 2 and positive sentiment, the positive correlation coefficient with `PFRV` might very likely stem from the missing consideration of negation phrases surrounding the positive words[55]. Also, in contrast to expectations, longer reports (i.e., larger `GFS`) on an univariate basis seem to correlate with lower `PFRV`, giving rise to the conjecture that the management obfuscation theory provided in F. Li (2008) might hold true in the available sample; i.e., the filing company might indeed succeed to disguise bad news by applying verbose language and producing long reports.

With regards to the control variables, all relationships except `TRVOL` and `LEVER` are as expected: smaller firms and firms with larger book-to-market ratio are, ceteris paribus, more risky. The latter, although significant, is however small in absolute magnitude. Surprisingly, and contradictory to previous findings, firms with large trading volume before the filing and a higher degree of leverage appear to co-occur with *lower* `PFRV` on a weekly horizon for the

[55] This alternative hypothesis seems to be supported by the high and significant positive correlation (.58) between `NEG_SENT` and `POS_SENT`.

sample at hand[56]. Moreover, market-wide volatility (VIX) seems to spill-over to company-level realized volatility post filing (correlation of .29 with PFRV).

6.2 Multivariate Analysis and Regression Results

In this section, the analysis will be extended to a multivariate setting and report OLS estimation results for equation (4.1). Overall, five annual augmented MZ-regressions (2013-2017) were estimated in three different scenarios, each one representing on of the three specifications for the sample splitting methodology (i.e., static, rolling, and extending VIBTW weight training window, respectively).

Table 8 reports the results for the static weight estimation window. Surprisingly, goodness-of-fit increases for regressions till 2015 and then starts to decrease again, although the weight estimation window is static. However, as sentiment scores are significant only in a few cases, this performance discrepancies are very likely to stem from the non-textual variables included in the linear model. As expected, the coefficient related to PreFRV forecast is highly significant and positive in the multivariate MZ-regression setting as well. With respect to the textual variables, negative sentiment increases PFRV significantly in all five out-of-sample years available for testing. For positive sentiment, the multivariate analysis confirms the findings from correlation analysis: post-filing realized volatility *increases* with POS_SENT, and this is - as for negative sentiment - the case for all five out-of-sample years. These results indicate that positive sentiment actually might be a hidden version of negative sentiment, with bad news being disguised by negated versions of otherwise positive phrases. However, with respect to the other hypotheses, assertiveness and uncertainty in the writing style of the 10-K* filing do not seem to affect PFRV, while litigious language seems to do so only in the latter period of the testing sample (2016 and 2017). Readability and intensity of financial terms are significantly different from zero only in single specifications (i.e., the MZ-regression in 2015 and 2013, respectively). With respect to readability, the positive coefficient attached to GFS thus at least for one year confirms hypothesis 6 (yet contradicts the univariate analysis) and implies - on ten percent significance level - that longer and thereby less readable reports increase realized volatility in the week after the filing. As indicated, and regarding financial focus in the text, hypothesis 7 is confirmed: usage of figures, ratios, percentages, monetary keywords, etc. decreases PFRV as investors perceive the news in the 10-K* as less informative. However, the statistical evidence being present

[56] Note that for TRVOL the result might be due to the timing difference related to the variable construction; while PFRV measures volatility *post*-filing, TVOL represents the median number of shares traded *pre*-filing. If any positive relationship between volume and volatility were to hold, it would likely occur if the variables are measured concurrently. Indeed, one needs to acknowledge that trading volume is very likely to behave differently pre- and post-filing, especially on a weekly horizon. However, for this thesis all independent variables were constructed in such fashion that they are *known* to the investor before the filing, making the volatility prediction in fact a forecasting task.

only in one out of five years available for testing weakens the relative importance of these variables, especially when compared to the strong significance of negative and (potentially false) positive tonality.

With respect to the control variables, one can conclude that large companies, firms with low financial leverage as well as firms with low book-to-market ratio are, ceteris paribus, perceived as less risky, implying lower PFRV. For trading volume the results are insignificant in the first testing year (2013) and then turn significantly positive for 2014-2017. For VIX, the results are ambiguous: for 2013, the estimation reveals the apparently paradoxical relation of decreasing firm-level volatility when markets are turbulent. For the middle part of the test sample at hand (2014 - 2016) no significant impact of market-level volatility on firm-specific PFRV is found, with the relation turning to the (expected) positive co-movement in 2017.

Table 9 shows the regression estimates for the rolling VIBTW estimation window, while Table 10 does so for the extending training window. The results are almost equivalent to the static window, indicating that the choice of the weight estimation window is not of crucial importance over the time horizon of multiple years. For all textual variables the results are close to identical for all specifications, i.e., significance as well as positive coefficients for negative and positive sentiment as well as punctual significance for assertive, uncertain and litigious language as well as readability and usage of financial terminology.

As an alternative to running five separate annual MZ-regression, one can also regress all out-of-sample observations in a single, pooled linear model. This specification is considered in robustness checks when different term weighting schemes are compared; the results are presented in Table 13. Anticipating the results presented therein, the coefficients associated with 10-K* language are both qualitatively and quantitatively similar to the estimates for annual regression presented in this section. In addition, Table 11 presents pooled OLS estimates that compare coefficient estimates after the addition / deletion of textual variables rather than across different weighting schemes. For what concerns positive, negative, and litigious language, the coefficients are significant and positive regardless of the inclusion of the other text variables. Especially worth-mentioning is the change in the coefficient of NEG_SENT after positive sentiment is included; indeed, the coefficient declines, indicating that POS_SENT is in fact a hidden version of negative tonality and "soaks up" variation in PFRV. A similar picture is obtained as soon as LITI is added, as coefficients of both positive and negative sentiment decline. This observation might point towards collinearity between those three variables, as all three might in fact be rephrased versions of "bad news". With regards to the usage of litigious sentiment words, it is necessary to highlight that significance in the pooled OLS estimates is likely to stem from the past two years in the test sample (i.e., 2016 and 2017), as those were the only years that corroborated the significant findings also in the five annual MZ regressions. Finally, the results of the stepwise addition of textual

coefficients also reveal that readability (GFS) is only significant when financial keyword intensity is not included in the model.

All in all it can be said that the results across different regression variants are very similar and point towards a common conclusion: when using past volatility, which was proven to be a good predictor for current/future volatility, in combination with common control variables such as VIX-level, the firm's size, leverage, and book-to-market ratio, or the trading volume of the stock, these variables already capture a majority of variation in post-filing realized volatility. Textual content of the 10-K* seems to improve predictive accuracy considerably and consistently only for the positive and negative dimensions of language. Other, "finer" facets of text (like assertiveness versus uncertainty, or readability) fall short in providing value added in predicting realized volatility in the week after the filing was submitted, or do so only for specific years of the test sample and/or a specific choice of the training window size. Although the overall goodness-of-fit of the different models would display potential for improvement, the latter does not seem to stem from the textual sentiment embedded in the filing. Two potential explanations for the insignificant results for the five text variables other than positive and negative tonality could be a poor performance of the VIBTW-scheme or a mis-measurement of the chosen volatility proxy. The next section is, among one other test performed, devoted to test result robustness by addressing exactly these two ideas.

7 Robustness Checks and Potential Alternative Designs

In this section I will report and comment on some robustness checks that were performed. As a first examination, the main contribution of this thesis (the introduction of term weighting based on past volatility) will be tested against benchmark term weighting schemes. The second subsection deals with the question whether predictive performance can be improved when different volatility proxies are applied instead of `PFRV`. The third and final subsection seeks to test the robustness of the results as well as changes in the importance of 10-K* language for post-filing realized volatility when one uses more potent variables to replace pre-filing realized volatility: thus, instead of `PreFRV`, two time-series model forecasts ((GJR-) GARCH) will be used in the multivariate analysis.

7.1 Tests Against Conventional Benchmark Term-Weighting Schemes

As already indicated when presenting the definition of conventional weighting schemes, the benchmarks against which `VIBTW` shall be tested will be the following: term frequency-inverse document frequency (`TFIDF`), relative frequency-inverse document frequency (`RFIDF`), one plus log-weighted frequencies with and without inverse document frequency (`WF_1PLOG` and `WFIDF_1PLOG`, respectively), log-one-plus weighted frequencies with and without inverse document frequency (`WF_LOG1P` and `WIFDF_LOG1P`, respectively), and maximum-scaled term frequencies (`TFMAX`).

Table 12 presents five correlation matrices, each representing the five textual variables (`NEG_SENT`, `POS_SENT`, `ASSERT`, `UNCERT`, and `LITI`) and each being of dimension eight by eight (i.e., one measure for each of the weighting schemes) and. It is evident from the Table that weighting schemes `WF_1PLOG` (`WFIDF_1PLOG`) and `WF_LOG1P` (`WIFDF_LOG1P`) correlate (almost) perfectly; thus, it seems irrelevant whether the log-transform is applied on raw counts and then added to one or whether counts are increased by unity and then logged. With the exception of assertiveness in the language (variable `ASSERT`), the volatility-based weighting scheme correlates positively (and in virtually all cases significantly) with the other scores, although the absolute magnitude is rather small (below .30 in all cases). As for other descriptive statistics involving `VIBTW`, the estimations are based on the full sample of 46,483 annual filings. Moreover, those scores that take account of inverse document frequency correlate heavier among each other compared to correlations with the log/max-weighted term frequencies; the latter, in turn, correlate among each other more intensely.

These observations give rise to estimate the linear model in (4.1) with alternative weighting schemes so as to benchmark market-impact-based weighting against established, conventional methods. In order to ensure presentability, regressions were estimated in a

pooled OLS model for years 2013-2017 for all eight weighting schemes[57]. The results are shown in Table 13. In general, there is no large difference in goodness-of-fit, yet the VIBTW scheme delivers the best performance. Speaking additionally in favour of volatility-based term weighting is the relative economic size of the estimated coefficients, which besides statistical significance implies that the appearance of words that co-occurred with high volatility in the past is likely to induce higher post-10-K*-filing realized volatility as well. Surprisingly, and indicating that weights on the basis of market reaction seem to smooth out the importance compared to other weighting designs, assertive language (ASSERT) is highly significant for the WF_1PLOG and WF_LOG1P schemes, respectively. The same argument holds true for the focus on financial topics (FIN) and the TFMAX weighting scheme. Another interesting fact that is revealed by Table 13 is that log-transformed (IDF) term weights disagree with the VIBTW scheme when it comes to the impact uncertain language in a 10-K* has on PFRV. While past-volatility-based weighting leads to a (expected) significant positive relation for LITI, columns (4) to (7) surprisingly indicate that language of uncertainty *decreases* post-filing realized volatility. Concerning the control variables, it is worth mentioning that – irrespective of the word weighting scheme applied in the textual analysis part – all results are significant and with expected sign. This is especially noteworthy for the level of market calmness/turbulence (measured by the level of the VIX), which displayed ambiguous results in the annual cross-sectional regressions.

7.2 Alternative Volatility Proxies

The second robustness check applied in this thesis relates to the left-hand side in both weight regressions and the augmented MZ-regression, and is thus connected with the measurement of post-filing volatility. As indicated in section 3.4, in this thesis a measure of weekly *realized* volatility was applied. Two commonly regarded alternative volatility proxies are squared returns (PFSqR)and absolute returns (PFAbsR). However, the regression results do not seem to be driven by measurement error in the dependent variable: as Table 6, which was presented in the section on results, indicates, the two alternative volatility measures correlate almost one-for-one with the realized volatility variable PFRV (correlation coefficients of .98 for PFSqR and .97 for PFAbsR, respectively).

Based on this strong similarity between the different measures, one does not expect perceptible changes when testing the models against alternative proxies of volatility. To examine the effect of how the volatility variable is computed, I estimated pooled regressions of the linear model in equation (4.1) with VIBTW weights coming from 1999-2012 for the two chosen alternative specifications, using either squared or absolute

[57] Therefore, both VIBTW and IDF-weights were estimated from 10-K* filings from 1999 to 2012.

returns as surrogate proxies. The results of these two pooled regressions are displayed in Table 14. Probably attributable to the usage of "noisier" post-filing volatility proxies, the overall goodness of fit measured by the model's R-squared is about five percentage points higher compared to the realized volatility approach. This manifests also in the jump of the size of β_1, the coefficient attributed to `PreFRV`, which increases from .039 in the realized volatility model to .299 for squared returns and .294 for absolute returns, respectively. The main findings related to 10-K* language are hardly distinguishable from the pooled regression using realized volatility, i.e., mainly negative and positive sentiment affect post-filing volatility. As for the model using `PFRV`, litigious language is also significant when using the pooled regression approach, whereas it was only relevant for several years when tested in the annual regression framework. Moreover, resembling prior results, the usage of `PFSqR` and `PFAbsR` does not alter the importance of textual assertiveness, readability and the degree of financial keywords; all of them are insignificant in providing incremental explanatory power. The appearance of uncertain language, however, is significant on ten percent level and increases post-filing volatility, if the latter is measured by squared or absolute returns in the week after the 10-K* submission. Finally, all estimates related to the control variables are almost identical to the ones obtained using the proxy `PFRV` (implying significance and expected sign in both specifications).

7.3 Substituting Pre-Filing Realized Volatility with Time-Series Inputs

The third and last robustness check to be performed concerns the replacement of the variable `PreFRV` with more powerful predictors of firm-specific volatility. Instead of using simply the "past observation", i.e., pre-filing realized volatility, as a linear input in a regression framework, I will use 1-week ahead forecasts from two widely-used time-series models as predictors. The models of choice are the GARCH and GJR-GARCH[58]. The corresponding models that were used to produce 1-week ahead forecasts were estimated on weekly return data from one year prior to the filing date (or, alternatively, the maximum number of weekly data available, if this number is smaller than 52). The two time-series models shall be used in order to test the idea whether a much more potent predictor still leaves room for volatility to be explained by quantitative information contained in the 10-K* filing, and, thus, affects the significance of the text-related variables. The latter notion stems from the idea that – compared to `PreFRV` – the forecasts from (GJR-) GARCH might "absorb" the most variation in `PFRV`, thereby leaving little room for sentiment variables to improve the model's performance.

[58] The main motivation for the choice of these two models is to have both one symmetric and one asymmetric model; GARCH(1,1) and GJR-GARCH(1,1) have proven to be parsimonious yet effective and hard-to-beat benchmarks (e.g., also considered in Antweiler and Frank (2004) and Hansen and Lunde (2005)).

The robustness check can be performed in eight variants (two pooled out-of-sample OLS regressions with static training window for both the GARCH- as well as the GJR-GARCH-model, or, alternatively, annual regressions using static/rolling/extending weight training windows for both GARCH and the GJR-GARCH). The regression results for the pooled regressions are presented in Tables 15 and 16. For the pooled regressions, the goodness-of-fit of the models is generally higher (about 8 percentage points larger R-Squared compared to the baseline that used pre-filing realized volatility). This is true for all specifications of weighting scheme. With respect to the textual variables within the `VIBTW` methodology (displayed in the first column of Tables 15 and 16), the regressions using (GJR-)GARCH partly confirm the conjecture that potent time-series models "soak up" the variation in `PFRV`. While both negative and positive tonality in the annual report remain significant, litigious language (`LITI`) is not (less) significant when using the GARCH (GJR-GARCH) inputs.

The annual MZ regressions presented in Tables 17 through 22 also in part confirm the idea of an significance "absorption" by the two powerful time-series models: considering that there are 35 coefficients of particular interest for each of the models (seven textual variables times five out-of-sample years), the baseline model using pre-filing realized volatility "relies" heavier on the textual inputs, with the number of significant entries out of the 35 possible in all cases exceeding those of the (GJR-) GARCH based estimators, with the exception of the rolling training window. Table 23 summarizes the number of significant textual variables for each of the models and training set methodologies (static, rolling, and extending) presented in this chapter.

7.4 Adding 10-K* Related Quantitative Control Variables

The last and most important robustness check concerns the isolation of textual content in the 10-K* from the simultaneous release of quantitative information contained therein. To illustrate this point, one can think of the following example: the fact that the filing company uses the (negatively connoted) word loss might per se **not** be driving volatility at all – and any impact from a negative sentiment score in such a case might be instead fully attributable to the fact that the company reported a loss in the first place, which is negative news for investors on its own. Hence, 10-K* language might be used merely to describe, accentuate, support, or alleviate the quantitative results of the fiscal year, without providing value added above any of the metrics reported in the same filing. Thus, controlling for the simultaneous release of investor-relevant metrics is necessary. A good orientation for this task is again Feldman et al. (2010), who controlled for simultaneous earnings, accruals, and cash-flow-based measures to isolate the effect of sentiment in the domain for equity returns.

Using data from Compustat (WRDS), I opted to include the following 10-K* related quantitative information:

- P&L turnaround (PnL_TURN): boolean that indicates if a company has managed to turn around a loss into a profit (or maintain a profit) between two consecutive filings
- Balance sheet growth (SIZE_CHG): relative change in total assets between two consecutive filings
- Change in book-to-market ratio between two consecutive filings (BTM_CHG)
- Relative change in the leverage ratio between two consecutive filings (LEVER_CHG)
- Sales growth (SALES_GROWTH): relative change in revenues between two consecutive filings
- Share buyback (SHR_BUYBACK): boolean that indicates whether the number of shares outstanding has been reduced between two consecutive filings
- Change in profitability (PROF_CHG): relative change in the profit margin

Note that two filings are considered to be "consecutive" if their filing year is not more than two years apart. Moreover, due to missing information for some entities, adding these variables reduces the size of the test set to 11,759 filings (previously: 13,679).

In addition to the seven variables presented above, I included two interaction variables: positive sentiment times P&L turnaround, labelled POS_SENTxPnL, and positive sentiment times change in profitability, labelled POS_SENTxPROF_CHG. These variables aim to control for a potential usage of false positive language. The idea is that firms might make excessive use of positive sentiment words whenever they have to convey negative news to investors (this could be either because they accentuate other positive things more heavily, or, use negation constructs like not increased to in fact describe decreased).

The results of the extended models is displayed in Tables 24 through 27. Table 27 displays the estimation in a pooled OLS setting, and - in comparison with 13 - shows that negative, positive and litigious language remain significant even after the introduction of the 10-K* related metrics. Similarly, firm size and leverage, book-to-market ratio, the level of the VIX and the trading volume are still significant predictors of PFRV when the additional quantitative control variables are included in the model. With respect to the newly variables it is worth noting that only the share buyback indicator and the change in leverage are significant on levels of 99 percent and 95 percent of confidence, respectively. The overall fit of the extended model is about four percentage points higher than in the standard version (R-Squared of 36.6% compared to 32.8%). However, inspecting Tables 24, 25, and 26 one can observe that the effect of including quantitative metrics is heterogeneous across years and model specification in terms of weight estimation method (static / rolling / extending). While in some cases the metric-effect seems to drive out the language-effect, in other constellations text-related variables become (or in most cases: remain) significant.

One last remark shall be made on the data used to construct the control variables. All figures are computed from balance sheet and profit-and-loss statements from the Compustat database and are always chosen to come from the end of year before the 10-K* was filed. In most cases these figures are filled/corrected in retrospective (for instance, the download of the revenues for a illustrative company X might read 50 million USD as per year-end 2017, i.e., 31/12/2017. However, very often this figure is revealed in the 10-K* filing released in, say, March 2018, and then retroactively filled to match the financial year 2017). As the subject of analysis in this work are 10-K* filings and these are released annually (with the exception of amendments), for all cases where the fundamental data for the previous financial year is available, these figures should therefore coincide with the metrics reported in the filing itself.

8 Conclusions

Being a variable of interest for both academia and practical application, volatility forecasting is a complex undergoing with many factors to consider. Besides classical time-series models, which – due to the auto-regressive nature of return volatility – appear to be the main driver of realized volatility even after corporate filing releases (such as 10-K* submissions), many researchers attempted to find innovative methods to include further explanatory variable that assist in explaining post-filing volatility. This thesis contributes to this particular stream of literature by analyzing textual contents in corporate filings. On the basis of Jegadeesh and Wu (2013), market-based term weighting of 10-K* text content was transferred to the field of volatility modelling: using past observations of realized volatility after the submission of a 10-K*, sentiment word counts are smoothed on the basis of the impact that specific word had on volatility. The learned smoothing factors, called volatility-impact based term weights (`VIBTW`) were then applied to out-of-sample term counts for "newly" filed 10-K* reports. The weighted term schemes were then – in an aggregated manner – used in augmented MZ regressions in order to test whether textual features such as tone or readability can provide incremental value to conventional models.

However, the main results of the estimations seem to resemble a well-known stylized fact in empirical finance: while equity returns still seem to unpredictable, return volatility – at least to some extent – can be explained (especially by deploying established time-series methods). In this line of thought, the equity return puzzle seems to be more "open" for improvements coming from the world of textual analysis (if term weighting is derived from past realizations of (abnormal) returns). Yet, based on sentiment word lists from Loughran and McDonald (2011) and Loughran and McDonald (2014) and a sample of 46,483 10-K* filings, this work finds a significant relationship between post-filing realized volatility for negative and positive tonality embedded in annual reports. Other text-related variables, such as assertiveness, uncertainty, litigiousness, focus on financial terminology, or readability seem to play an insignificant role in the augmented MZ regressions which were estimated – therefore leaving the majority of variance in realized volatility in the cross section to be explained by known factors. Regarding the latter, a robustness check that takes use of two models firmly established in the volatility literature, revealed that the by far most important model inputs are time-series model forecasts such as the GARCH and the GJR-extension thereof. These factors are accompanied by established control variables, such as the level of the VIX, firm size, leverage, trading volume, and book-to-market ratio. These findings are robust in terms of how one chooses to proxy unobservable post-filing volatility: squared and absolute returns confirm the results that were obtained using realized volatility, implying that the construction of the volatility variable does not seem to play a role when determining the impact of quantitative and qualitative inputs: Similarly, the findings are also robust

with regards to the choice of the term-weighting scheme. In great part, the results derived from a market-based term smoothing (`VIBTW`) are confirmed by established methodologies such as `TFIDF` or `WFIDF`, or are even superseded by the newly introduced `VIBTW`.

However, although this thesis contributing to an ever growing body of literature of text analysis in finance and accounting, it is worth highlighting that this field of research is still "young" and leaves room for improvement. The most obvious extension is that future contributions might consider corpora that are not exclusively in English language. Especially for the business domain, foreign languages still lag behind with the presence of extensive dictionaries or sentiment word lists, thereby leaving large room for prospective contributions.

Moreover, with respect to the textual underlying used for sentiment analysis in this work, 10-K* present one of the most important inputs for text analysis in the finance and accounting domain. At the same time, one needs to acknowledge one key shortcoming of this corpus, which – unfortunately – can not be circumvented: any annual report is a description about a company's *past* fiscal year. Although there are requirements which force the management to include forward-looking statements in the 10-K* as well (such as the mentioned Item 7, i.e., the so-called MD&A section), a large part of text is still reflecting past occurrences and results. This ultimately raises the question whether the tone extracted from such a filing is suitable for the attempt to explain market variables in the future.

Recognizing this unavoidable disadvantage related to the 10-K* corpus, there are, however, potential improvements to the methodologies applied in this thesis. Hence, I will provide some impulses for future research and present potential alternative research specifications. A first and feasible starting point regards the choice of the post-event window (labelled τ in equation (3.12)). In other words, instead of considering realized volatility in the *week* after the 10-K* filing date, variations might apply monthly (e.g. Jegadeesh and Wu (2013)), quarterly (e.g., Rekabsaz et al. (2017)) or annual (e.g., Loughran and McDonald (2011)) volatility measures instead.

Another promising specification might be to follow Nizer and Nievola (2012), who"revert" the idea presented in this thesis: instead of using classified text content (i.e., sentiment) to explain volatility, the authors propose that "abnormal" volatility might be a means of its own that helps to categorize news (or in the context of this work: a 10-K*) as important/informative. Abnormality of volatility is measured by the forecast error: "a way to evaluate the news' importance is by the error (difference) between a volatility model and the effective volatility" (Nizer and Nievola 2012, p. 10675). In the project at hand this could be defined as a delta series of the post- and pre-filing realized volatility (i.e., `PFRV` minus `PreFRV`). Also the forecasts derived from the (GJR-)GARCH model could be used to compute the relevant error series. As another potential variant, one might also define

abnormality as a post-filing volatility level that deviates from an average level of volatility observed in a pre-defined past time horizon (such as one year).

Furthermore, in the field of volatility prediction one often makes use of concepts like absolute returns or squared returns, which in turn implies that the *direction* of the stock price movement per se does not matter. This leaves the additional possibility for future research to conduct sentiment analysis in a modified fashion: instead of categorizing sentiment into "buckets" such as negative, positive, uncertain, etc. tone, one might alternatively realize that it is not necessarily the type of sentiment that influences post-filing volatility, but rather the sheer presence of sentiment (of any kind). In other words, corporate filings in this context need not be classified as positive, negative, assertive, and so on – it would be sufficient to find a measure that classifies the textual content as stock price *relevant* versus stock price *irrelevant*. In such a setting, one would "merely" care whether the textual sentiment of an annual report will affect the stock return, regardless in which direction the price change occurs. However, this would call for a modified and cautious definition of the document's "sentiment" score, which in this case might be instead called "relevance" score: ignoring context and simply summing up sentiment word counts in such a setting *might* lead to the effect of **cancelling out** of tone-opposing frequencies (e.g., one positive and one negative statement could potentially be either two relevant entries (unconnected statements) or instead zero relevant entries (with, for instance, the positive statement attenuating the relevance of the negative statement – or vice versa)).

Yet another potential extension of the presented analysis is to consider *changes* in sentiment instead of *levels*, that is conduct research on the basis of first differences in sentiment scores (or term counts). In this context, Loughran and McDonald (2016, p. 1220) point out: "when document tone is differenced, the importance of Zipf's law is mitigated. Zipf's law applies to levels, but the distribution of changes in tone can be driven by numerous words that are not the most frequently occurring overall words". However, as Loughran and McDonald (2011) indicated, "The differencing method also assumes that a reader can remember the frequency of negative words in previous news articles, columns, or 10-Ks – for example, that today's column or 10-K has fewer negative words than previous editions, so it may convey a bullish signal." A good starting point for analysis pointing in this direction are the papers of Loughran and McDonald (2011) and Feldman et al. (2010).

References

Amel-Zadeh, Amir and Jonathan Faasse (2016). The Information Content of 10-K Narratives: Comparing MD&A and Footnotes Disclosures. Working Paper, available under `https://www.wiwi.hu-berlin.de/de/professuren/bwl/cofi/seminars/16w/amel-zadeh_2016.pdf` (visited on 05/27/2018) (cit. on pp. 14, 32).

Antweiler, Werner and Murray Z. F. Frank (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance* 59.3, pp. 1259–1294 (cit. on pp. 2, 11, 30, 33, 49).

Belsky, Gary (2012). *Why Text Mining May Be The Next Big Thing*. TIME. URL: `http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/` (visited on 05/19/2018) (cit. on p. 3).

Bollerslev, Tim (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31.3, pp. 307–327 (cit. on p. 1).

Brownlees, Christian, Robert F. Engle, and Bryan T. Kelly (2011). A Practical Guide to Volatility Forecasting through Calm and Storm. *Journal of Risk* 14.2, pp. 3–22 (cit. on p. 27).

Das, Sanjiv R. (2014). Text and Context: Language Analytics in Finance. *Foundations and Trends® in Finance* 8.3, pp. 145–261 (cit. on pp. 2–3, 12).

De Franco, Gus, Ole-Kristian Hope, Dushyantkumar Vyas, and Yibin Zhou (2015). Analyst Report Readability. *Contemporary Accounting Research* 32.1, pp. 76–104 (cit. on pp. 8–9).

Diebold, Francis X. (2017). Forecasting in Economics, Business, Finance and Beyond. Department of Economics, University of Pennsylvania. Available under `http://www.ssc.upenn.edu/~fdiebold/Textbooks.html` (visited on 07/24/2018) (cit. on p. 35).

Engle, Robert F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50.4, pp. 987–1007 (cit. on p. 1).

Engle, Robert F. and Andrew J. Patton (2001). What good is a volatility model? *Quantitative Finance* 1, pp. 237–245 (cit. on p. 27).

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15.4, pp. 915–953 (cit. on pp. 13, 50, 55).

Franck, Thomas (2018). *Obscure security linked to stock volatility plummets 80% after hours, sparking worries of bigger market effect*. CNBC. URL: `https://www.cnbc.com/2018/`

02/05/xiv-exchange-traded-security-linked-to-volatility-plummets-80-percent.html (visited on 03/18/2018) (cit. on p. 2).

Gries, Stefan T. (2009). Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge (Taylor & Francis Group) (cit. on p. 18).

Guo, Li, Feng Shi, and Jun Tu (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science* 2.3, pp. 153–170 (cit. on p. 6).

Hansen, Peter R. and Asger Lunde (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20.7, pp. 873–889 (cit. on p. 49).

Heidari, Maryam and Carsten Felden (2015). Impact of Text Mining Application on Financial Footnotes Analysis. *New Horizons in Design Science: Broadening the Research Agenda*. Ed. by Brian Donnellan, Markus Helfert, Jim Kenneally, Debra VanderMeer, Marcus Rothenberger, and Robert Winter. Springer International Publishing, pp. 463–470 (cit. on p. 14).

Heires, Katherine (2015). *Sentiment Analysis: Are You Feeling Risky?* Risk Management Magazine. URL: http://www.rmmagazine.com/2015/12/01/sentiment-analysis-are-you-feeling-risky/ (visited on 04/24/2018) (cit. on pp. 3–4).

Hsieh, Chia C., Kai W. Hui, and Yao Zhang (2016). Analyst Report Readability, Earnings Uncertainty and Stock Returns. *Journal of Business Finance & Accounting* 43 (1), pp. 98–130 (cit. on p. 9).

Huang, Allen H., Amy Y. Zang, and Rong Zheng (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review* 89.6, pp. 2151–2180 (cit. on pp. 11, 31–33).

Huang, Ke-Wei and Zhuolun Li (2008). A Multilabel Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K. *ACM Transactions on Management Information Systems (TMIS)* 2.3, 18:1–18:19 (cit. on p. 13).

Jegadeesh, Narasimhan and Di Wu (2013). Word power: A new approach for content analysis. *Journal of Financial Economics* 110.3. Accepted Manuscript, available under https://www.sciencedirect.com/science/article/pii/S0304405X13002328 (visited on 05/21/2018), pp. 712–729 (cit. on pp. 15, 19, 25–27, 53–54).

Jones, Charles M., Gautam Kaul, and Marc L. Lipson (1994). Transactions, Volume, and Volatility. *Review of Financial Studies* 7.4, pp. 631–651 (cit. on p. 33).

Jurafsky, Dan and James H. Martin (2017). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Draft of the 3rd edition - preprint as of 08/28/2017, available under https://web.stanford.edu/~jurafsky/slp3/ (visited on 05/11/2018) (cit. on pp. 15–19).

Kearney, Colm and Sha Liu (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33, pp. 171–185 (cit. on pp. 6, 12).

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith (2009). Predicting Risk from Financial Reports with Regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, USA: Association for Computational Linguistics, pp. 272–280 (cit. on pp. 12–13, 29, 40, 42).

Kothari, S. P., Xu Li, and James E. Short (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review* 84.5, pp. 1639–1670 (cit. on pp. 14, 30, 33).

Kremer, Andreas, Florian Strobel, and Wolfgang Malzkorn (2013). *Ratings revisited: Textual analysis for better risk management*. McKinsey & Company. URL: https://www.mckinsey.com/business-functions/risk/our-insights/ratings-revisited-textual-analysis-for-better-risk-management (visited on 04/24/2018) (cit. on pp. 3–4, 10).

Kumar, Shravan and Vadlamani Ravi (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems* 114, pp. 128–147 (cit. on pp. 2–3).

Lehavy, Reuven, Feng Li, and Kenneth Merkley (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review* 86.3, pp. 1087–1115 (cit. on pp. 9, 14).

Li, Feng (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45.2, pp. 221–247 (cit. on pp. 8, 32, 43).

Li, Feng (2010). The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research* 48.5, pp. 1049–1102 (cit. on p. 13).

Loughran, Tim and Bill McDonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66.1, pp. 35–65 (cit. on pp. 11, 14–15, 27, 30–31, 38, 53–55).

Loughran, Tim and Bill McDonald (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance* 69.4, pp. 1643–1671 (cit. on pp. 9–10, 14, 32, 38, 53).

Loughran, Tim and Bill McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research* 54.4, pp. 1187–1230 (cit. on pp. 2–3, 6–8, 10, 13–14, 18, 21, 30, 32, 55).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). Introduction to Information Retrieval. Cambridge University Press (cit. on pp. 16, 21–22, 24).

Markowitz, Harry (1952). Portfolio Selection. *The Journal of Finance* 7.1, pp. 77–91 (cit. on p. 1).

Mittnik, Stefan, Nikolay Robinzonov, and Martin Spindler (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking & Finance* 58, pp. 1–14 (cit. on p. 2).

Nizer, Philippe S.M. and Julio Cesar Nievola (2012). Predicting published news effect in the Brazilian stock market. *Expert Systems with Applications* 39.12, pp. 10674–10680 (cit. on p. 54).

Patton, Andrew J. and Kevin Sheppard (2009). Evaluating Volatility and Correlation Forecasts. *Handbook of Financial Time Series*. Ed. by Thomas Mikosch, Jens-Peter Kreiß, Richard A. Davis, and Torben Gustav Andersen. Springer Berlin Heidelberg, pp. 801–838 (cit. on pp. 27–28, 35).

Paye, Bradley S. (2012). 'Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics* 106.3, pp. 527–546 (cit. on p. 2).

Pulliza, Jonathan L. (2015). An Analysis of Speculative Language in SEC 10-K Filings. MA thesis. University of North Carolina at Chapel Hill (cit. on pp. 13–14).

Qiu, Xin Ying (2007). On building predictive models with company annual reports. PhD thesis. University of Iowa (cit. on p. 14).

Raviv, Eran (2012). *Intraday volatility measures*. URL: `https://eranraviv.com/intraday-volatility-measures/` (visited on 05/21/2018) (cit. on p. 28).

Rekabsaz, Navid, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, and Linda Anderson (2017). Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Vancouver, Canada, pp. 1712–1721 (cit. on pp. 12–13, 18, 22, 29, 40, 54).

Schutte, Maria G. and Emre Unlu (2007). The Effect of Analyst Coverage on Firm-Specific Volatility: Less Information or Less Noise? *Financial Analyst Journal* (cit. on p. 9).

Tetlock, Paul C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62.3, pp. 1139–1168 (cit. on pp. 11, 33).

Thinggaard, Frank, Carsten Sønderby Jeppesen, and Kasper Madsen (2015). The Information Content of Note Disclosures and MD&A Information in the Financial Report – A Study of Market Reactions in Denmark. *Ledelse & Erhvervsoekonomi* 79.4, pp. 25–42 (cit. on p. 14).

Tsai, Ming-Feng and Chuan-Ju Wang (2012). Visualization on Financial Terms via Risk Ranking from Financial Reports. *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India: International Conference on Computational Linguistics (COLING), pp. 447–452 (cit. on pp. 12–13, 42).

Tsai, Ming-Feng and Chuan-Ju Wang (2013). Risk Ranking from Financial Reports. *Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR) - LNCS7814*. Ed. by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz. Moscow, Russia: Springer, pp. 804–807 (cit. on pp. 12–13, 42).

Tsai, Ming-Feng and Chuan-Ju Wang (2014). Financial Keyword Expansion via Continuous Word Vector Representations. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar (cit. on pp. 12, 18, 22).

Tsai, Ming-Feng and Chuan-Ju Wang (2016). On the Risk Prediction and Analysis of Soft Information in Finance Reports. *European Journal of Operational Research* 257.1, pp. 243–250 (cit. on pp. 12–13, 18, 22, 29, 42).

Tsai, Ming-Feng, Chuan-Ju Wang, and Po-Chuan Chien (2016). Discovering Finance Keywords via Continuous-Space Language Models. *ACM Transactions on Management Information Systems* 7.3 (cit. on pp. 12–13).

Varian, Hal R. (2004). *Good Stock Advice or Online Noise?* The New York Times. URL: `https://www.nytimes.com/2004/09/23/business/good-stock-advice-or-online-noise.html` (visited on 04/20/2018) (cit. on p. 2).

Violante, Francesco and Sébastien Laurent (2012). Volatility Forecasts Evaluation and Comparison. *Handbook of Volatility Models and Their Applications*. Wiley-Blackwell. Chap. Nineteen, pp. 465–486 (cit. on p. 35).

Wang, Chuan-Ju, Tse Tsai Ming-Fengand Liu, and Chin-Ting Chang (2013). Financial Sentiment Analysis for Risk Prediction. *International Joint Conference on Natural Language Processing (IJCNLP)*. Nagoya, Japan, pp. 802–808 (cit. on pp. 12–13).

Yukselturk, Osman and Jon Tucker (2015). The impact of analyst sentiment on UK stock recommendations and target prices. *Accounting and Business Research* 45, pp. 869–904 (cit. on pp. 6, 11).

Zobel, Justin and Alistair Moffat (1998). Exploring the Similarity Space. *ACM SIGIR Forum* 32.1, pp. 18–34 (cit. on p. 24).

Tables

Table 1: Sample Cleaning and Matching Procedure

Description	Number of filings	Number of CIKs	Filings per CIK	Percentage of 1^{st} row‡
Initial sample of 10-K*s:	295,746	40,264	7.3	100.0%
Ticker available:	184,358	12,525	14.7	62.3%
Yahoo! Finance data available:	77,951	4,934	15.8	26.4%
COMPUSTAT data available:	64,527	4,125	15.6	21.8%
Data meaningful†:	46,483	3,736	12.4	15.7%

†: Meaningful refers to a number of criteria:first off, all data points that are filled with the WRDS placeholder `NA` instead of numbers are excluded. Additionally, for `PFRV`, `PreFRV`, `SIZE`, `BTM`, and `TRVOL` only strictly positive values are included in the sample (so as to allow operating in the log-space further on). Finally, all observations with a financial leverage ratio above one (implying negative book value of equity) are dropped.

‡: The percentage refers to the number of *filings* remaining in the sample, i.e., the second column of this table.

Table 2: Sample Distribution: Filing Types

10-K	10-K-A	10-K405	10-K405-A	10-KSB	10-KSB-A	10-KT
36,659	6,103	1,872	252	1	1	56
10-KT-A	10KSB	10KSB-A	10KSB40	10KSB40-A	10KT405	
16	1,041	381	75	21	5	

Table 3: Corpus Composition and Size by Years

Year	# of filings	Total number of tokens	Total number of types	Mean number of tokens	Mean number of types	Median number of tokens	Median number of types
1999	1,808	81 MM	7 MM	44,612	3,775	31,964	3,643
2000	1,968	87 MM	7 MM	44,360	3,771	31,713	3,633
2001	2,059	98 MM	8 MM	47,543	3,901	34,477	3,777
2002	1,956	97 MM	8 MM	49,368	3,914	36,704	3,855
2003	2,141	117 MM	9 MM	54,470	4,193	40,951	4,076
2004	2,302	123 MM	10 MM	53,407	4,162	41,942	4,127
2005	2,626	135 MM	11 MM	51,494	4,049	42,072	4,069
2006	2,490	135 MM	11 MM	54,017	4,240	44,445	4,174
2007	2,426	135 MM	11 MM	55,499	4,358	47,164	4,316
2008	2,480	143 MM	11 MM	57,796	4,414	48,784	4,398
2009	2,547	161 MM	12 MM	63,369	4,530	52,840	4,524
2010	2,515	153 MM	11 MM	60,827	4,523	51,831	4,539
2011	2,687	166 MM	12 MM	61,886	4,597	53,480	4,580
2012	2,799	171 MM	13 MM	60,972	4,541	53,960	4,614
2013	2,787	174 MM	13 MM	62,354	4,621	54,998	4,652
2014	2,771	182 MM	13 MM	65,822	4,769	57,391	4,768
2015	2,773	180 MM	13 MM	64,934	4,745	57,680	4,757
2016	2,687	179 MM	13 MM	66,678	4,825	59,237	4,816
2017	2,661	186 MM	13 MM	70,071	4,918	61,139	4,910
Total	46,483	2,703 MM	204 MM	58,147	4,399	49,117	4,415

Table 4: Top-5-Words in the Corpus by Sentiment Category

Full Sample		**1999-2004**		**2005-2010**		**2011-2017**	
count	word	count	word	count	word	count	word
			Negative				
5.18	cost	1.01	will	1.70	cost	2.55	cost
4.90	will	0.94	cost	1.57	will	2.31	will
3.28	loss	0.55	loss	1.00	loss	1.74	loss
2.51	subject	0.51	termin	0.76	subject	1.28	subject
2.44	limit	0.50	limit	0.75	limit	1.19	limit
			Positive				
4.43	effect	0.87	effect	1.52	effect	2.04	effect
3.13	inform	0.57	inform	1.00	inform	1.56	inform
2.75	benefit	0.55	benefit	0.96	benefit	1.24	benefit
1.23	posit	0.31	except	0.42	posit	0.61	posit
1.14	except	0.21	posit	0.36	except	0.47	except
			Uncertainty				
7.49	may	1.40	may	2.26	may	3.83	may
2.78	could	0.53	approxim	0.87	approxim	1.68	could
2.63	condit	0.46	condit	0.81	condit	1.35	condit
2.46	approxim	0.33	risk	0.80	could	1.34	risk
2.38	risk	0.30	could	0.71	risk	1.07	approxim
			Litigious				
6.45	shall	2.36	shall	2.02	shall	2.31	will
4.90	will	1.01	will	1.57	will	2.08	shall
2.36	amend	0.58	amend	0.76	amend	1.06	record
2.22	record	0.51	termin	0.76	record	1.02	amend
2.02	contract	0.45	law	0.66	contract	1.00	regul
			Constraining				
5.47	requir	1.04	requir	1.73	requir	2.70	requir
2.44	limit	0.51	oblig	0.75	limit	1.19	limit
2.18	oblig	0.50	limit	0.69	oblig	0.98	oblig
1.51	impair	0.26	direct	0.47	restrict	0.86	impair
1.51	restrict	0.24	restrict	0.47	impair	0.80	restrict
			Strong Modal				
4.90	will	1.01	will	1.57	will	2.31	will
0.60	must	0.12	must	0.19	must	0.29	must
0.44	definit	0.10	definit	0.13	definit	0.21	definit
0.28	best	0.06	best	0.09	best	0.13	best
0.15	strong	0.03	strong	0.05	strong	0.07	strong
			Modest Modal				
7.49	may	1.40	may	2.26	may	3.83	may
2.78	could	0.30	could	0.80	could	1.68	could
0.74	depend	0.11	depend	0.23	depend	0.40	depend
0.40	possibl	0.07	possibl	0.12	possibl	0.21	possibl
0.20	appear	0.05	appear	0.07	appear	0.11	uncertain
			Weak Modal				
2.89	general	0.56	general	0.93	general	1.40	general
1.71	would	0.33	would	0.54	would	0.84	would
1.14	can	0.19	can	0.36	can	0.59	can
0.60	like	0.09	should	0.20	like	0.34	like
0.53	should	0.07	like	0.17	should	0.26	should

Notes: All numbers are in millions. All counts are based on raw term frequencies. Wherever the counts of the three subperiods do not sum up to the total sample counts, the discrepancies are attributable to rounding.

Table 5: Sample Summary: Variable Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
PFRV	46,483	−3.82	.86	−8.76	−4.38	−3.86	−3.30	2.39
PreFRV	46,483	−3.87	.85	−8.73	−4.41	−3.90	−3.37	1.97
NEG_SENT	46,483	.01	.01	−.01	.01	.01	.02	.06
POS_SENT	46,483	−.11	.02	−.22	−.13	−.11	−.10	.03
ASSERT	46,483	−.16	.02	−.36	−.17	−.16	−.15	.24
UNCERT	46,483	−.02	.01	−.10	−.03	−.02	−.01	.17
LITI	46,483	.03	.001	.02	.03	.03	.03	.04
GFS	46,483	14.55	1.77	7.14	13.29	14.46	16.11	19.89
FIN	46,483	5.87	1.11	0.00	5.67	6.11	6.48	8.12
SIZE	46,483	6.40	2.32	−6.91	4.85	6.51	7.93	14.73
BTM	46,483	−.82	.96	−10.68	−1.27	−.71	−.23	8.30
TRVOL	46,483	11.60	2.58	0.00	9.90	11.89	13.44	20.05
VIX	46,483	19.68	8.37	9.43	13.84	17.86	23.18	72.67
LEVER	46,483	.52	.25	0.00	.32	.52	.72	1.00

Notes: As indicated in the description of the variable construction (sections 3.4, 4.2, and 4.3, respectively), the variables PFRV, PreFRV, GFS, FIN, SIZE, BTM, and TRVOL are expressed in logarithmic form. All one-hot encoded boolean variables (i.e., YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are omitted from this table.

Table 6: Correlation Matrix: Volatility-Related Measures

	PFRV	PreFRV	GARCH	GJR	PFSqR	PFAbsR
PFRV	1.00	.00	.00	.00	.00	.00
PreFRV	.58	1.00	.00	.00	.00	.00
GARCH	.65	.68	1.00	.00	.00	.00
GJR_GARCH	.60	.65	.90	1.00	.00	.00
PFSqR	.98	.60	.66	.62	1.00	.00
PFAbsR	.97	.59	.65	.61	.99	1.00

Notes: Sample Size = 46,483. Pearson correlation coefficients are displayed in the lower triangular part (including the main diagonal) of the matrix; the upper triangular part represents the corresponding p-values for the correlation coefficients. All correlation coefficients as well as p-values are based on volatility-related variables in their logarithmic form. `PFRV` represents post-filing realized volatility. `PreFRV` represents pre-filing realized volatility. `GARCH` and `GJR_GARCH` represent the 1-week ahead volatility forecast from a GARCH(1,1) and a GJR-GARCH(1,1) model, respectively (cf. robustness checks on time-series models, section 7.3). `PFSqR` and `PFAbsR` represent post-filing squared returns and post-filing absolute returns, respectively (cf. robustness checks on alternative volatility proxies, section 7.2).

Table 7: Correlation Matrix: All Variables

	PFRV	PreFRV	NEG_SENT	POS_SENT	ASSERT	UNCERT	LITI	GFS	FIN	SIZE	BTM	TRVOL	VIX	LEVER
PFRV	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
PreFRV	.58	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
NEG_SENT	.44	.41	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
POS_SENT	.33	.32	.58	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.26	.00
ASSERT	.19	.17	.34	.41	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
UNCERT	.32	.31	.56	.56	.34	1.00	.00	.00	.00	.00	.00	.00	.00	.00
LITI	.35	.33	.60	.57	.35	.54	1.00	.00	.00	.00	.00	.00	.00	.00
GFS	-.20	-.21	-.30	-.16	-.07	-.30	-.31	1.00	.00	.00	.00	.00	.00	.00
FIN	-.10	-.11	-.18	-.18	-.05	-.24	-.20	.56	1.00	.00	.00	.00	.30	.00
SIZE	-.38	-.36	-.59	-.58	-.31	-.55	-.54	.34	.30	1.00	.00	.00	.00	.00
BTM	.01	.01	-.09	-.17	-.10	-.17	-.12	.02	.03	.14	1.00	.00	.00	.03
TRVOL	-.16	-.13	-.24	-.19	-.06	-.15	-.19	.27	.19	.62	-.21	1.00	.00	.00
VIX	.29	.31	.12	.01	.02	.08	.09	-.27	.00	-.02	.10	-.02	1.00	.37
LEVER	-.16	-.15	-.37	-.41	-.26	-.45	-.39	.15	.18	.46	-.01	.03	.00	1.00

Notes: Sample Size = 46,483. Pearson correlation coefficients are displayed in the lower triangular part (including the main diagonal) of the matrix; the upper triangular part represents the corresponding p-values for the correlation coefficients. Textual variables, where applicable, were constructed using the VIBTW scheme, with term weights learned from the full sample (1999 - 2017). All one-hot encoded boolean variables (i.e., YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are omitted from this table.

Table 8: Annual Augmented MZ-Regression (Static Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
PreFRV	.064***	.052***	−.003	.023***	.049***
	(.008)	(.009)	(.008)	(.008)	(.011)
NEG_SENT	7.771***	6.194**	8.713***	5.260**	3.761*
	(2.702)	(2.439)	(2.258)	(2.277)	(2.187)
POS_SENT	4.516***	2.956**	2.291**	5.012***	2.808**
	(1.150)	(1.174)	(1.069)	(1.072)	(1.104)
ASSERT	.047	.582	.653	−.512	.111
	(.438)	(.424)	(.404)	(.438)	(.469)
UNCERT	2.232*	1.282	−.506	1.260	−.882
	(1.265)	(1.133)	(1.218)	(1.274)	(1.451)
LITI	14.385	22.906	22.794	41.234***	33.581**
	(14.725)	(13.993)	(14.129)	(13.677)	(14.562)
GFS	−.008	.029	.042**	.036	−.011
	(.020)	(.023)	(.020)	(.024)	(.026)
FIN	.043*	.016	−.033	−.018	.012
	(.022)	(.025)	(.021)	(.025)	(.028)
SIZE	−.141***	−.166***	−.200***	−.208***	−.227***
	(.013)	(.013)	(.012)	(.013)	(.013)
BTM	.083***	.072***	.090***	.183***	.185***
	(.017)	(.017)	(.016)	(.016)	(.017)
TRVOL	.016*	.032***	.067***	.120***	.087***
	(.010)	(.009)	(.009)	(.010)	(.011)
VIX	−.057***	.027**	.007	.018	.063**
	(.017)	(.014)	(.008)	(.011)	(.029)
LEVER	.397***	.246***	.538***	.654***	.660***
	(.079)	(.076)	(.074)	(.074)	(.080)
Constant	−3.102***	−3.695***	−.117	−6.408***	−5.545***
	(.911)	(1.067)	(.994)	(.989)	(.972)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.328	.361	.410	.365	.363
Adjusted R^2	.296	.331	.382	.334	.332

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. All VIBTW weights were estimated using 10-K* filings from 1999 to 2012. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 9: Annual Augmented MZ-Regression (Rolling Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
PreFRV	.064***	.054***	−.003	.024***	.052***
	(.008)	(.009)	(.008)	(.009)	(.011)
NEG_SENT	7.929***	−.205	3.084	−11.798**	−7.188
	(2.700)	(7.267)	(5.796)	(5.189)	(5.189)
POS_SENT	4.529***	3.816	−2.602	−19.786	.674
	(1.149)	(3.270)	(3.279)	(17.677)	(5.086)
ASSERT	.028	−.315	−1.068	−1.135	−1.279
	(.437)	(1.555)	(1.647)	(1.182)	(2.151)
UNCERT	2.172*	.849	−1.608	4.551*	−.245
	(1.264)	(1.631)	(1.530)	(2.386)	(2.558)
LITI	14.716	4.205	3.900	15.380	35.351*
	(14.713)	(9.586)	(21.481)	(17.288)	(18.175)
GFS	−.008	.013	.031	.025	−.015
	(.020)	(.023)	(.020)	(.024)	(.026)
FIN	.044**	.040	−.014	−.011	.021
	(.022)	(.025)	(.021)	(.025)	(.027)
SIZE	−.140***	−.202***	−.235***	−.252***	−.254***
	(.013)	(.011)	(.011)	(.012)	(.012)
BTM	.081***	.074***	.092***	.187***	.190***
	(.017)	(.017)	(.016)	(.016)	(.017)
TRVOL	.015	.042***	.079***	.135***	.098***
	(.010)	(.009)	(.009)	(.010)	(.010)
VIX	−.057***	.029**	.005	.016	.062**
	(.017)	(.014)	(.008)	(.011)	(.029)
LEVER	.396***	.192**	.508***	.633***	.649***
	(.079)	(.076)	(.074)	(.074)	(.080)
Constant	−3.117***	−3.216***	1.436	−5.952***	−4.757***
	(.911)	(.975)	(1.773)	(1.567)	(1.005)
Observations	2,785	2,771	2,771	2,687	2,660
R^2	.328	.353	.398	.353	.358
Adjusted R^2	.297	.322	.370	.321	.326

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. VIBTW weights were estimated using 10-K* filings from the past 14 years (i.e., (Y-14) to (Y-1)). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 10: Annual Augmented MZ-Regression (Extending Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
PreFRV	.064***	.052***	−.003	.022***	.049***
	(.008)	(.009)	(.008)	(.008)	(.011)
NEG_SENT	7.929***	13.444***	14.611***	10.601***	18.694***
	(2.700)	(5.169)	(3.430)	(3.675)	(3.988)
POS_SENT	4.529***	2.973***	2.425**	6.883***	1.713*
	(1.149)	(1.103)	(.990)	(1.804)	(.986)
ASSERT	.028	.681	.814	−.627	−.078
	(.437)	(.466)	(.507)	(.634)	(.724)
UNCERT	2.172*	1.144	.141	2.203	−.017
	(1.264)	(1.138)	(1.288)	(1.459)	(1.684)
LITI	14.716	21.363	9.805	40.102***	43.294***
	(14.713)	(13.477)	(13.771)	(13.412)	(15.299)
GFS	−.008	.028	.042**	.036	−.017
	(.020)	(.023)	(.020)	(.024)	(.026)
FIN	.044**	.018	−.024	−.009	.037
	(.022)	(.025)	(.021)	(.025)	(.027)
SIZE	−.140***	−.164***	−.193***	−.202***	−.206***
	(.013)	(.013)	(.012)	(.013)	(.014)
BTM	.081***	.072***	.089***	.184***	.177***
	(.017)	(.017)	(.016)	(.016)	(.017)
TRVOL	.015	.032***	.065***	.120***	.081***
	(.010)	(.009)	(.009)	(.010)	(.011)
VIX	−.057***	.027**	.007	.019*	.055*
	(.017)	(.014)	(.008)	(.011)	(.029)
LEVER	.396***	.246***	.532***	.658***	.653***
	(.079)	(.076)	(.074)	(.073)	(.079)
Constant	−3.117***	−3.692***	.314	−6.882***	−5.773***
	(.911)	(1.048)	(.976)	(.966)	(.964)
Observations	2,785	2,771	2,771	2,687	2,660
R^2	.328	.362	.409	.366	.371
Adjusted R^2	.297	.332	.381	.335	.340

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. VIBTW weights were estimated using 10-K* filings from 1999 to (Y-1). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 11: Augmented MZ-Regression: Stepwise Addition / Deletion of Textual Variables

	Dependent variable: PFRV						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PreFRV	.041***	.040***	.040***	.039***	.040***	.039***	.039***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
NEG_SENT	9.545***	7.223***	7.012***	6.699***	5.995***	6.022***	6.011***
	(.990)	(1.017)	(1.027)	(1.041)	(1.053)	(1.053)	(1.053)
POS_SENT		4.421***	4.252***	4.083***	3.691***	3.704***	3.692***
		(.467)	(.481)	(.490)	(.498)	(.498)	(.498)
ASSERT			.283	.267	.191	.147	.133
			(.192)	(.192)	(.193)	(.193)	(.194)
UNCERT				1.001*	.796	.916	.930*
				(.555)	(.557)	(.559)	(.559)
LITI					27.090***	27.443***	27.215***
					(6.341)	(6.342)	(6.349)
GFS						.019**	.015
						(.008)	(.010)
FIN							.008
							(.011)
SIZE	−.202***	−.186***	−.186***	−.184***	−.182***	−.185***	−.185***
	(.005)	(.005)	(.005)	(.006)	(.006)	(.006)	(.006)
BTM	.119***	.119***	.120***	.120***	.120***	.119***	.119***
	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)
TRVOL	.067***	.061***	.061***	.061***	.060***	.060***	.060***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
VIX	.022***	.023***	.023***	.023***	.023***	.023***	.023***
	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
LEVER	.458***	.477***	.477***	.484***	.487***	.479***	.477***
	(.034)	(.034)	(.034)	(.034)	(.034)	(.034)	(.034)
Constant	−3.414***	−2.966***	−2.937***	−2.905***	−3.958***	−4.223***	−4.190***
	(.304)	(.306)	(.307)	(.308)	(.394)	(.410)	(.412)
Observations	13,679	13,679	13,679	13,679	13,679	13,679	13,679
R^2	.322	.327	.327	.327	.328	.328	.328
Adjusted R^2	.316	.320	.321	.321	.322	.322	.322

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Textual variables, where applicable, were constructed using the VIBTW scheme, with the correpsonding weights being estimated from a fixed training set (1999-2012).

Table 12: Correlation Matrix: Scores Calculated by Different Term-Weighting Schemes

	VIBTW	TFIDF	RFIDF	WFIDF_ 1PLOG	WFIDF_ LOG1P	WF_1PLOG	WF_LOG1P	TFMAX
				NEG_SENT				
VIBTW	1.00	.00	.01	.00	.00	.00	.00	.00
TFIDF	.06	1.00	.00	.00	.00	.00	.00	.00
RFIDF	.01	.10	1.00	.00	.00	.00	.00	.00
WFIDF_1PLOG	.20	.65	.28	1.00	.00	.00	.00	.00
WFIDF_LOG1P	.20	.67	.28	1.00	1.00	.00	.00	.00
WF_1PLOG	.17	-.32	.28	.22	.20	1.00	.00	.00
WF_1PLOG	.17	-.32	.28	.22	.21	1.00	1.00	.00
TFMAX	.02	-.23	.48	-.06	-.06	.59	.59	1.00
				POS_SENT				
VIBTW	1.00	.00	.37	.00	.00	.00	.00	.01
TFIDF	.26	1.00	.00	.00	.00	.00	.00	.00
RFIDF	.00	.16	1.00	.00	.00	.00	.00	.00
WFIDF_1PLOG	.25	.59	.38	1.00	.00	.00	.00	1.00
WFIDF_LOG1P	.26	.61	.38	1.00	1.00	.00	.00	.80
WF_1PLOG	.14	-.16	.16	.25	.24	1.00	.00	.00
WF_1PLOG	.14	-.16	.16	.25	.24	1.00	1.00	.00
TFMAX	.01	-.13	.46	.00	.00	.61	.61	1.00
				ASSERT				
VIBTW	1.00	.00	.00	.00	.00	.00	.00	.00
TFIDF	-.26	1.00	.00	.00	.00	.00	.00	.00
RFIDF	-.12	.28	1.00	.00	.00	.00	.00	.00
WFIDF_1PLOG	-.15	.81	.40	1.00	.00	.00	.00	.00
WFIDF_LOG1P	-.16	.82	.40	1.00	1.00	.00	.00	.00
WF_1PLOG	-.18	-.06	.19	.10	.10	1.00	.00	.00
WF_1PLOG	-.18	-.06	.19	.10	.10	1.00	1.00	.00
TFMAX	-.09	-.18	.25	-.08	-.08	.87	.87	1.00
				UNCERT				
VIBTW	1.00	.14	.00	.00	.00	.00	.00	.00
TFIDF	.01	1.00	.00	.00	.00	.00	.00	.00
RFIDF	.06	.11	1.00	.00	.00	.00	.00	.00
WFIDF_1PLOG	.13	.79	.34	1.00	.00	.00	.00	.00
WFIDF_LOG1P	.13	.80	.34	1.00	1.00	.00	.00	.00
WF_1PLOG	.14	-.18	.47	.22	.21	1.00	.00	.00
WF_1PLOG	.14	-.17	.47	.22	.22	1.00	1.00	.00
TFMAX	.06	-.22	.62	-.04	-.04	.65	.65	1.00
				LITI				
VIBTW	1.00	.36	.00	.00	.00	.00	.00	.00
TFIDF	.00	1.00	.00	.00	.00	.00	.00	.00
RFIDF	.07	-.02	1.00	.00	.00	.00	.00	.00
WFIDF_1PLOG	.11	.44	.31	1.00	.00	.00	.00	.00
WFIDF_LOG1P	.11	.47	.31	1.00	1.00	.00	.00	.00
WF_1PLOG	.14	-.47	.55	.14	.11	1.00	.00	.00
WF_1PLOG	.14	-.48	.55	.14	.11	1.00	1.00	.00
TFMAX	.06	-.21	.35	-.11	-.12	.51	.50	1.00

Notes: Sample Size = 46,483. Pearson correlation coefficients are displayed in the lower triangular part (including the main diagonal) of the matrix; the upper triangular part represents the corresponding p-values for the correlation coefficients.

Table 13: Augmented MZ-Regression: Pooled OLS

	Dependent variable: PFRV							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PreFRV	.039***	.041***	.041***	.041***	.041***	.040***	.040***	.041***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
NEG_SENT	6.011***	.391***	2.861	.381***	.563***	.030	.049	−.003
	(1.053)	(.092)	(1.985)	(.148)	(.217)	(.176)	(.264)	(.003)
POS_SENT	3.692***	.414***	4.143	−.013	.022	.460***	.715***	.013*
	(.498)	(.088)	(3.487)	(.221)	(.324)	(.153)	(.227)	(.008)
ASSERT	.133	.245	−3.558	−.021	−.036	.360***	.494***	.053
	(.194)	(.179)	(7.822)	(.304)	(.449)	(.124)	(.184)	(.048)
UNCERT	.930*	−.156	−.382	−.339	−.515	−.195	−.332*	.0002
	(.559)	(.179)	(1.335)	(.268)	(.393)	(.134)	(.197)	(.003)
LITI	27.215***	.068	4.867	−.520**	−.672*	−1.125***	−1.643***	.010
	(6.349)	(.065)	(17.650)	(.260)	(.378)	(.164)	(.241)	(.023)
GFS	.015	.005	.002	.004	.004	−.001	−.001	.003
	(.010)	(.010)	(.010)	(.010)	(.010)	(.010)	(.010)	(.010)
FIN	.008	.003	.031**	.026**	.026**	.014	.011	.045***
	(.011)	(.011)	(.012)	(.011)	(.011)	(.016)	(.016)	(.014)
SIZE	−.185***	−.220***	−.223***	−.222***	−.222***	−.224***	−.224***	−.224***
	(.006)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)
BTM	.119***	.127***	.123***	.123***	.123***	.122***	.122***	.122***
	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)
TRVOL	.060***	.069***	.072***	.072***	.072***	.069***	.069***	.072***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
VIX	.023***	.022***	.022***	.022***	.022***	.022***	.022***	.022***
	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
LEVER	.477***	.468***	.441***	.447***	.448***	.432***	.434***	.435***
	(.034)	(.034)	(.034)	(.034)	(.034)	(.034)	(.034)	(.034)
Constant	−4.190***	−3.379***	−3.289***	−3.270***	−3.280***	−3.009***	−2.979***	−3.396***
	(.412)	(.324)	(.328)	(.328)	(.328)	(.347)	(.348)	(.330)
Observations	13,679	13,679	13,679	13,679	13,679	13,679	13,679	13,679
R^2	.328	.321	.318	.318	.318	.321	.321	.318
Adjusted R^2	.322	.314	.312	.312	.312	.314	.314	.312

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table. Column headers (1) through (8) refer to the weighting schemes in the following order: VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, WFIDF_LOG1P, WF_1PLOG, WF_LOG1P, TFMAX. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Weights for the VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, and WFIDF_LOG1P are estimated from a fixed training set (1999-2012).

Table 14: Robustness Check: Pooled OLS with Alternative Volatility Proxies

	Dependent variable:	
	Squared Returns (PFSqR)	Absolute Returns (PFAbsR)
PreFRV	.299***	.294***
	(.008)	(.008)
NEG_SENT	4.357***	4.502***
	(.979)	(1.010)
POS_SENT	3.687***	4.028***
	(.548)	(.606)
ASSERT	.118	.143
	(.184)	(.176)
UNCERT	.912*	.955*
	(.490)	(.496)
LITI	17.135***	12.837**
	(5.750)	(5.215)
GFS	.010	.010
	(.009)	(.009)
FIN	.014	.017*
	(.010)	(.010)
SIZE	−.126***	−.129***
	(.005)	(.005)
BTM	.088***	.090***
	(.007)	(.007)
TRVOL	.034***	.047***
	(.004)	(.004)
VIX	.015***	.016***
	(.003)	(.003)
LEVER	.352***	.348***
	(.032)	(.031)
Constant	−2.353***	−1.829***
	(.360)	(.331)
Observations	13,679	13,679
R^2	.399	.394
Adjusted R^2	.394	.388

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. Textual variable scores were calculated using VIBTW, for which the weights were estimated using all available 10-K* filings from 1999 to 2012. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 15: Robustness Check: Substituting Pre-Filing Realized Volatilty with GARCH Forecasts

	Dependent variable: PFRV							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GARCH	.596***	.609***	.613***	.613***	.613***	.610***	.610***	.613***
	(.013)	(.013)	(.013)	(.013)	(.013)	(.013)	(.013)	(.013)
NEG_SENT	2.576***	.065	.084	.083	.122	−.023	−.029	−.003
	(.989)	(.086)	(1.841)	(.137)	(.202)	(.164)	(.246)	(.003)
POS_SENT	1.927***	.161**	4.196	.019	.030	.273*	.431**	.002
	(.468)	(.082)	(3.245)	(.205)	(.302)	(.143)	(.211)	(.007)
ASSERT	.012	.327*	8.033	.337	.494	.403***	.590***	.051
	(.182)	(.167)	(7.282)	(.283)	(.418)	(.116)	(.172)	(.045)
UNCERT	−.042	.035	.274	−.004	−.015	−.039	−.074	.002
	(.524)	(.167)	(1.242)	(.250)	(.366)	(.125)	(.183)	(.003)
LITI	9.311	.079	24.453	−.250	−.306	−.644***	−.972***	.019
	(5.957)	(.060)	(16.411)	(.242)	(.352)	(.153)	(.225)	(.022)
GFS	.009	.004	.005	.005	.005	.004	.004	.006
	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)
FIN	.002	−.002	.020*	.008	.008	.028*	.027*	.030**
	(.010)	(.010)	(.011)	(.010)	(.010)	(.015)	(.015)	(.013)
SIZE	−.088***	−.101***	−.101***	−.101***	−.100***	−.104***	−.104***	−.102***
	(.006)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)
BTM	.063***	.065***	.063***	.063***	.063***	.062***	.062***	.062***
	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)
TRVOL	.011***	.014***	.015***	.014***	.015***	.014***	.014***	.014***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
VIX	.019***	.019***	.019***	.019***	.019***	.019***	.019***	.019***
	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
LEVER	.217***	.210***	.196***	.201***	.201***	.186***	.187***	.190***
	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)	(.032)
Constant	−2.256***	−1.960***	−2.022***	−1.925***	−1.930***	−2.011***	−2.001***	−2.056***
	(.389)	(.304)	(.306)	(.307)	(.307)	(.324)	(.325)	(.309)
Observations	13,679	13,679	13,679	13,679	13,679	13,679	13,679	13,679
R^2	.411	.410	.409	.409	.409	.410	.410	.410
Adjusted R^2	.405	.404	.404	.404	.404	.405	.405	.404

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GARCH is the 1-week ahead forecast from a GARCH(1,1) model. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table. Column headers (1) through (8) refer to the weighting schemes in the following order: VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, WFIDF_LOG1P, WF_1PLOG, WF_LOG1P, TFMAX. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Weights for the VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, and WFIDF_LOG1P are estimated from a static training set (1999-2012).

Table 16: Robustness Check: Substituting Pre-Filing Realized Volatilty with GJR-GARCH Forecasts

	Dependent variable: PFRV							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GJR_GARCH	.450***	.462***	.467***	.467***	.467***	.463***	.463***	.467***
	(.012)	(.012)	(.011)	(.011)	(.011)	(.012)	(.012)	(.011)
NEG_SENT	3.277***	.159*	−.190	.136	.206	−.037	−.046	−.003
	(1.006)	(.088)	(1.876)	(.140)	(.205)	(.167)	(.251)	(.003)
POS_SENT	2.448***	.252***	5.380	.165	.255	.309**	.492**	.005
	(.476)	(.083)	(3.305)	(.209)	(.308)	(.145)	(.215)	(.007)
ASSERT	.021	.250	3.872	.071	.103	.370***	.540***	.057
	(.185)	(.170)	(7.416)	(.288)	(.426)	(.118)	(.175)	(.045)
UNCERT	.222	−.011	.171	−.044	−.083	−.063	−.128	.001
	(.533)	(.170)	(1.265)	(.254)	(.373)	(.127)	(.187)	(.003)
LITI	15.151**	.077	20.858	−.289	−.358	−.755***	−1.131***	.013
	(6.055)	(.061)	(16.716)	(.246)	(.358)	(.156)	(.230)	(.022)
GFS	.010	.004	.004	.004	.004	.002	.002	.005
	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.010)
FIN	.005	−.0001	.023**	.013	.013	.021	.020	.031**
	(.010)	(.010)	(.012)	(.010)	(.010)	(.015)	(.015)	(.013)
SIZE	−.114***	−.131***	−.132***	−.131***	−.131***	−.135***	−.135***	−.133***
	(.006)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)
BTM	.073***	.077***	.074***	.074***	.074***	.073***	.073***	.073***
	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)	(.007)
TRVOL	.023***	.027***	.028***	.028***	.028***	.027***	.027***	.028***
	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)
VIX	.021***	.021***	.021***	.021***	.021***	.021***	.021***	.021***
	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
LEVER	.276***	.268***	.249***	.256***	.256***	.240***	.241***	.244***
	(.033)	(.033)	(.033)	(.033)	(.033)	(.033)	(.033)	(.033)
Constant	−2.780***	−2.309***	−2.318***	−2.260***	−2.267***	−2.209***	−2.195***	−2.356***
	(.395)	(.309)	(.312)	(.312)	(.312)	(.330)	(.331)	(.314)
Observations	13,679	13,679	13,679	13,679	13,679	13,679	13,679	13,679
R^2	.390	.388	.387	.387	.387	.388	.388	.387
Adjusted R^2	.385	.382	.381	.381	.381	.383	.383	.381

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GJR_GARCH is the 1-week ahead forecast from a GJR-GARCH(1,1) model. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table. Column headers (1) through (8) refer to the weighting schemes in the following order: VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, WFIDF_LOG1P, WF_1PLOG, WF_LOG1P, TFMAX. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Weights for the VIBTW, TFIDF, RFIDF, WFIDF_1PLOG, and WFIDF_LOG1P are estimated from a static training set (1999-2012).

Table 17: Annual Augmented MZ-Regressions: Results for GARCH + Static Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GARCH	.604***	.606***	.572***	.611***	.561***
	(.029)	(.031)	(.029)	(.031)	(.033)
NEG_SENT	4.932*	3.281	5.295**	1.052	1.551
	(2.539)	(2.303)	(2.114)	(2.134)	(2.077)
POS_SENT	2.748**	1.026	.435	3.019***	1.609
	(1.082)	(1.111)	(1.002)	(1.005)	(1.050)
ASSERT	−.062	.349	.462	−.518	−.077
	(.410)	(.400)	(.377)	(.408)	(.446)
UNCERT	.654	.232	−1.402	.488	−1.449
	(1.190)	(1.069)	(1.137)	(1.188)	(1.378)
LITI	−12.594	3.018	3.197	19.289	31.486**
	(13.883)	(13.191)	(13.205)	(12.797)	(13.836)
GFS	−.015	.002	.035*	.036	−.001
	(.019)	(.021)	(.019)	(.022)	(.025)
FIN	.031	.034	−.044**	−.028	.012
	(.021)	(.024)	(.020)	(.024)	(.026)
SIZE	−.054***	−.071***	−.103***	−.102***	−.133***
	(.013)	(.013)	(.012)	(.013)	(.014)
BTM	.042***	.033**	.045***	.091***	.117***
	(.016)	(.016)	(.015)	(.016)	(.017)
TRVOL	−.016*	−.009	.016*	.047***	.030***
	(.009)	(.009)	(.009)	(.010)	(.011)
VIX	−.051***	.033**	.008	.026**	.048*
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.189**	.027	.291***	.281***	.421***
	(.075)	(.073)	(.070)	(.071)	(.077)
Constant	−.966	−1.521	−.105	−3.938***	−3.672***
	(.862)	(1.011)	(.927)	(.930)	(.930)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.408	.433	.487	.448	.425
Adjusted R^2	.381	.406	.462	.421	.397

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GARCH is the 1-week ahead GARCH(1,1) forecast. All VIBTW weights were estimated using 10-K* filings from 1999 to 2012. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 18: Annual Augmented MZ-Regressions: Results for GJR-GARCH + Static Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GJR_GARCH	.500***	.434***	.416***	.439***	.413***
	(.026)	(.027)	(.025)	(.027)	(.028)
NEG_SENT	5.658**	4.168*	5.524**	2.499	1.804
	(2.565)	(2.350)	(2.154)	(2.179)	(2.104)
POS_SENT	3.102***	1.720	.985	3.621***	2.074*
	(1.092)	(1.132)	(1.019)	(1.027)	(1.063)
ASSERT	−.054	.310	.588	−.527	−.067
	(.414)	(.408)	(.384)	(.418)	(.452)
UNCERT	.444	.800	−1.200	.673	−.784
	(1.203)	(1.090)	(1.158)	(1.216)	(1.396)
LITI	−5.687	11.901	9.150	24.295*	33.864**
	(14.006)	(13.448)	(13.440)	(13.089)	(14.016)
GFS	−.022	.009	.040**	.041*	.001
	(.019)	(.022)	(.019)	(.023)	(.025)
FIN	.037*	.026	−.039*	−.027	.008
	(.021)	(.024)	(.020)	(.024)	(.027)
SIZE	−.072***	−.098***	−.132***	−.135***	−.157***
	(.013)	(.013)	(.012)	(.013)	(.014)
BTM	.043***	.043***	.055***	.113***	.135***
	(.016)	(.016)	(.016)	(.016)	(.017)
TRVOL	−.010	.002	.028***	.069***	.043***
	(.009)	(.009)	(.009)	(.010)	(.011)
VIX	−.046***	.027**	.007	.022**	.061**
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.208***	.081	.378***	.381***	.475***
	(.075)	(.074)	(.071)	(.072)	(.078)
Constant	−1.510*	−1.939*	−.017	−4.526***	−4.526***
	(.869)	(1.032)	(.944)	(.951)	(.938)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.396	.409	.467	.422	.410
Adjusted R^2	.367	.381	.442	.394	.381

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GJR_GARCH is the 1-week ahead GJR-GARCH(1,1) forecast. All VIBTW weights were estimated using 10-K* filings from 1999 to 2012. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 19: Annual Augmented MZ-Regressions: Results for GARCH + Rolling Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GARCH	.604***	.622***	.586***	.634***	.574***
	(.029)	(.031)	(.028)	(.030)	(.032)
NEG_SENT	4.932*	.463	2.660	−11.578**	−10.001**
	(2.539)	(6.806)	(5.376)	(4.797)	(4.915)
POS_SENT	2.748**	1.976	−.868	−16.481	1.072
	(1.082)	(3.064)	(3.041)	(16.339)	(4.818)
ASSERT	−.062	.276	−1.441	−.397	−1.198
	(.410)	(1.457)	(1.528)	(1.093)	(2.037)
UNCERT	.654	.730	−1.949	4.046*	.533
	(1.190)	(1.527)	(1.419)	(2.206)	(2.422)
LITI	−12.594	6.775	−2.523	15.766	44.260**
	(13.883)	(8.977)	(19.910)	(15.979)	(17.217)
GFS	−.015	−.004	.033*	.030	−.004
	(.019)	(.021)	(.018)	(.022)	(.025)
FIN	.031	.044*	−.036*	−.027	.020
	(.021)	(.024)	(.020)	(.024)	(.026)
SIZE	−.054***	−.081***	−.112***	−.116***	−.145***
	(.013)	(.012)	(.012)	(.013)	(.013)
BTM	.042***	.033**	.046***	.089***	.119***
	(.016)	(.016)	(.015)	(.016)	(.017)
TRVOL	−.016*	−.006	.018**	.051***	.034***
	(.009)	(.009)	(.009)	(.010)	(.010)
VIX	−.051***	.033**	.008	.025**	.048*
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.189**	.005	.284***	.263***	.418***
	(.075)	(.072)	(.069)	(.070)	(.077)
Constant	−.966	−1.624*	−.005	−3.924***	−2.889***
	(.862)	(.917)	(1.644)	(1.451)	(.957)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.408	.432	.485	.447	.425
Adjusted R^2	.381	.405	.461	.420	.396

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GARCH is the 1-week ahead GARCH(1,1) forecast. VIBTW weights were estimated using 10-K* filings from the past 14 years (i.e., (Y-14) to (Y-1)). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 20: Annual Augmented MZ-Regressions: Results for GJR-GARCH + Rolling Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GJR_GARCH	.500***	.451***	.431***	.462***	.424***
	(.026)	(.027)	(.024)	(.027)	(.027)
NEG_SENT	5.658**	−.163	4.004	−11.017**	−9.651*
	(2.565)	(6.959)	(5.484)	(4.919)	(4.984)
POS_SENT	3.102***	2.172	−1.077	−18.411	1.050
	(1.092)	(3.133)	(3.101)	(16.755)	(4.886)
ASSERT	−.054	.346	−1.360	−.451	−1.062
	(.414)	(1.490)	(1.558)	(1.121)	(2.066)
UNCERT	.444	.375	−1.473	3.818*	−.043
	(1.203)	(1.562)	(1.447)	(2.262)	(2.456)
LITI	−5.687	6.050	3.166	11.910	41.663**
	(14.006)	(9.178)	(20.304)	(16.388)	(17.457)
GFS	−.022	−.001	.036*	.034	−.003
	(.019)	(.022)	(.019)	(.023)	(.025)
FIN	.037*	.040*	−.028	−.025	.017
	(.021)	(.024)	(.020)	(.024)	(.026)
SIZE	−.072***	−.116***	−.147***	−.156***	−.174***
	(.013)	(.012)	(.011)	(.012)	(.013)
BTM	.043***	.043***	.055***	.111***	.137***
	(.016)	(.016)	(.016)	(.016)	(.017)
TRVOL	−.010	.007	.033***	.075***	.049***
	(.009)	(.009)	(.009)	(.010)	(.010)
VIX	−.046***	.027**	.006	.021*	.062**
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.208***	.044	.365***	.358***	.466***
	(.075)	(.073)	(.070)	(.072)	(.077)
Constant	−1.510*	−1.625*	.667	−4.600***	−3.692***
	(.869)	(.939)	(1.676)	(1.487)	(.967)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.396	.406	.464	.418	.408
Adjusted R^2	.367	.378	.439	.390	.379

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GJR_GARCH is the 1-week ahead GJR-GARCH(1,1) forecast. VIBTW weights were estimated using 10-K* filings from the past 14 years (i.e., (Y-14) to (Y-1)). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 21: Annual Augmented MZ-Regressions: Results for GARCH + Extending Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GARCH	.604***	.605***	.569***	.608***	.547***
	(.029)	(.031)	(.029)	(.031)	(.033)
NEG_SENT	4.932*	6.795	8.757***	3.315	11.630***
	(2.539)	(4.883)	(3.218)	(3.450)	(3.807)
POS_SENT	2.748**	1.213	.737	4.069**	.819
	(1.082)	(1.044)	(.929)	(1.689)	(.941)
ASSERT	−.062	.407	.535	−.693	−.283
	(.410)	(.439)	(.473)	(.592)	(.690)
UNCERT	.654	−.013	−.819	1.038	−1.152
	(1.190)	(1.074)	(1.203)	(1.362)	(1.606)
LITI	−12.594	4.092	−7.381	18.325	41.117***
	(13.883)	(12.702)	(12.884)	(12.558)	(14.576)
GFS	−.015	.002	.035*	.036	−.007
	(.019)	(.021)	(.019)	(.022)	(.025)
FIN	.031	.034	−.040**	−.024	.029
	(.021)	(.024)	(.020)	(.024)	(.026)
SIZE	−.054***	−.070***	−.100***	−.100***	−.122***
	(.013)	(.013)	(.012)	(.013)	(.014)
BTM	.042***	.034**	.045***	.091***	.114***
	(.016)	(.016)	(.015)	(.016)	(.017)
TRVOL	−.016*	−.009	.014	.048***	.027***
	(.009)	(.009)	(.009)	(.010)	(.011)
VIX	−.051***	.033**	.008	.026**	.044
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.189**	.028	.295***	.285***	.426***
	(.075)	(.073)	(.070)	(.071)	(.077)
Constant	−.966	−1.607	.221	−4.218***	−3.971***
	(.862)	(.993)	(.912)	(.912)	(.925)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.408	.433	.487	.448	.429
Adjusted R^2	.381	.407	.462	.421	.400

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GARCH is the 1-week ahead GARCH(1,1) forecast. VIBTW weights were estimated using 10-K* filings from 1999 to (Y-1). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 22: Annual Augmented MZ-Regressions: Results for GJR-GARCH + Extending Training Set

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
GJR_GARCH	.500***	.433***	.414***	.436***	.400***
	(.026)	(.027)	(.025)	(.027)	(.028)
NEG_SENT	5.658**	8.825*	9.752***	5.964*	12.264***
	(2.565)	(4.981)	(3.277)	(3.521)	(3.855)
POS_SENT	3.102***	1.913*	1.250	5.070***	1.173
	(1.092)	(1.063)	(.944)	(1.726)	(.952)
ASSERT	−.054	.368	.682	−.683	−.269
	(.414)	(.448)	(.482)	(.605)	(.699)
UNCERT	.444	.586	−.749	1.179	.069
	(1.203)	(1.096)	(1.226)	(1.394)	(1.624)
LITI	−5.687	11.260	−1.683	23.679*	43.558***
	(14.006)	(12.953)	(13.110)	(12.837)	(14.756)
GFS	−.022	.008	.040**	.041*	−.003
	(.019)	(.022)	(.019)	(.023)	(.025)
FIN	.037*	.027	−.035*	−.022	.027
	(.021)	(.024)	(.020)	(.024)	(.026)
SIZE	−.072***	−.097***	−.128***	−.132***	−.143***
	(.013)	(.013)	(.012)	(.013)	(.014)
BTM	.043***	.043***	.055***	.114***	.131***
	(.016)	(.016)	(.016)	(.016)	(.017)
TRVOL	−.010	.002	.027***	.069***	.040***
	(.009)	(.009)	(.009)	(.010)	(.011)
VIX	−.046***	.027**	.007	.023**	.057**
	(.016)	(.013)	(.007)	(.011)	(.028)
LEVER	.208***	.082	.381***	.385***	.481***
	(.075)	(.074)	(.071)	(.072)	(.077)
Constant	−1.510*	−1.949*	.323	−4.883***	−4.830***
	(.869)	(1.014)	(.929)	(.931)	(.932)
Observations	2,787	2,771	2,773	2,687	2,661
R^2	.396	.410	.468	.422	.414
Adjusted R^2	.367	.382	.442	.394	.385

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. GJR_GARCH is the 1-week ahead GJR-GARCH(1,1) forecast. VIBTW weights were estimated using 10-K* filings from 1999 to (Y-1). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 23: Significance of Textual Variables Across Baseline and Robustness Check Models

	Annual MZ Regressions with Traning Set Being ...			
	Static	Rolling	Extending	Pooled OLS
PreFRV	14	7	15	3
GARCH	7	9	8	2
GJR_GARCH	12	9	12	3

Notes: The table displays the number of significant textual variables (NEG_SENT, POS_SENT, ASSERT, UNCERT, LITI, GFS, and FIN) for each of the respective row-column model specifications. The maximum number for the three columns related to annual MZ regressions is 35 (seven textual variables times five out-of-sample test regressions); for the last column (pooled OLS) the maximum count is seven. In the pooled OLS model, textual variable significance is based on the model specification using the VIBTW scheme, with term weights estimated from 1999 to 2012.

Table 24: Annual Augmented MZ-Regression With Additional Filing-Related Control Variables (Static Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
TS_FC	.362***	.252***	.206***	.236***	.154***
	(.021)	(.021)	(.021)	(.021)	(.021)
NEG_SENT	3.656	1.513	8.068***	3.359	2.950
	(2.780)	(2.611)	(2.375)	(2.316)	(2.186)
POS_SENT	1.607	1.031	1.726	4.026***	2.557*
	(1.335)	(1.388)	(1.303)	(1.312)	(1.307)
ASSERT	.215	.239	.482	−.567	−.212
	(.437)	(.430)	(.415)	(.442)	(.471)
UNCERT	.267	.903	−1.611	1.067	−1.230
	(1.344)	(1.137)	(1.233)	(1.277)	(1.467)
LITI	7.827	5.573	2.102	37.681***	27.301*
	(15.292)	(14.492)	(14.634)	(13.929)	(14.638)
GFS	−.020	.018	.023	.044*	−.005
	(.022)	(.025)	(.022)	(.024)	(.027)
FIN	.044*	.014	−.019	−.030	.009
	(.023)	(.027)	(.023)	(.025)	(.028)
SIZE	−.088***	−.131***	−.133***	−.129***	−.183***
	(.016)	(.015)	(.014)	(.015)	(.016)
BTM	.072***	.077***	.068***	.114***	.154***
	(.021)	(.020)	(.019)	(.019)	(.020)
TRVOL	.005	.027***	.041***	.070***	.057***
	(.011)	(.010)	(.011)	(.011)	(.012)
VIX	−.086***	.019	−.001	.016	.049*
	(.018)	(.014)	(.008)	(.011)	(.030)
LEVER	.223**	.171*	.331***	.434***	.456***
	(.089)	(.088)	(.081)	(.079)	(.086)
PnL_TURN	.263	.181	.167	.049	.062
	(.162)	(.167)	(.154)	(.144)	(.156)
SIZE_CHG	−.050	.087	.076	.010	.049
	(.066)	(.059)	(.056)	(.055)	(.052)
BTM_CHG	.019	−.065**	.034	.054*	−.025
	(.035)	(.032)	(.032)	(.033)	(.034)
LEVER_CHG	−.016	−.004*	−.009	.004	−.008
	(.015)	(.003)	(.017)	(.016)	(.026)

SALES_GROWTH	−.001	.008	−.003	.001	.001
	(.001)	(.009)	(.002)	(.0005)	(.002)
SHR_BUYBACK	−.021	−.098***	−.084***	−.053*	−.094***
	(.034)	(.034)	(.032)	(.031)	(.032)
PROF_CHG	.016	−.002	.003	.023***	−.002
	(.011)	(.005)	(.009)	(.009)	(.002)
POS_SENTxPnL	1.435	.119	.322	−1.328	−1.460
	(1.952)	(2.061)	(1.899)	(1.731)	(1.857)
POS_SENTxPROF_CHG	.189	−.036	.062	.422***	−.038
	(.133)	(.066)	(.113)	(.160)	(.037)
Constant	−1.500	−1.901	.239	−5.283***	−4.963***
	(1.402)	(1.449)	(1.384)	(1.622)	(1.461)
Observations	2,308	2,277	2,362	2,405	2,407
R^2	.396	.397	.436	.404	.393
Adjusted R^2	.359	.359	.402	.369	.357

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. All VIBTW weights were estimated using 10-K* filings from 1999 to 2012. Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 25: Annual Augmented MZ-Regression With Additional Filing-Related Control Variables (Rolling Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
TS_FC	.362***	.255***	.210***	.244***	.158***
	(.021)	(.021)	(.021)	(.021)	(.021)
NEG_SENT	3.656	−1.297	3.288	−9.154*	−10.137*
	(2.780)	(7.396)	(5.849)	(5.215)	(5.240)
POS_SENT	1.607	3.922	−2.948	−11.444	.706
	(1.335)	(3.908)	(3.854)	(20.841)	(6.149)
ASSERT	.215	1.272	−2.091	−.718	−1.865
	(.437)	(1.609)	(1.655)	(1.183)	(2.166)
UNCERT	.267	.105	−1.121	4.133*	.484
	(1.344)	(1.653)	(1.591)	(2.373)	(2.542)
LITI	7.827	12.687	3.243	18.614	24.668
	(15.292)	(9.572)	(21.810)	(17.803)	(18.202)
GFS	−.020	.011	.017	.035	−.006
	(.022)	(.024)	(.021)	(.024)	(.027)
FIN	.044*	.022	−.007	−.027	.013
	(.023)	(.027)	(.023)	(.026)	(.028)
SIZE	−.088***	−.141***	−.150***	−.153***	−.198***
	(.016)	(.014)	(.014)	(.014)	(.015)
BTM	.072***	.078***	.071***	.113***	.157***
	(.021)	(.020)	(.019)	(.019)	(.020)
TRVOL	.005	.029***	.049***	.078***	.063***
	(.011)	(.010)	(.011)	(.011)	(.011)
VIX	−.086***	.019	−.002	.016	.049*
	(.018)	(.014)	(.008)	(.011)	(.030)
LEVER	.223**	.145*	.307***	.391***	.444***
	(.089)	(.087)	(.080)	(.079)	(.085)
PnL_TURN	.263	.572	−1.580**	−1.119	.927
	(.162)	(.443)	(.769)	(2.756)	(1.020)
SIZE_CHG	−.050	.100*	.096*	.018	.060
	(.066)	(.059)	(.056)	(.055)	(.052)
BTM_CHG	.019	−.067**	.032	.053	−.021
	(.035)	(.032)	(.032)	(.033)	(.034)
LEVER_CHG	−.016	−.004*	−.009	.008	−.004
	(.015)	(.003)	(.017)	(.016)	(.026)

SALES_GROWTH	−.001	.008	−.002	.001	.002
	(.001)	(.009)	(.002)	(.0005)	(.002)
SHR_BUYBACK	−.021	−.101***	−.096***	−.062**	−.099***
	(.034)	(.034)	(.032)	(.031)	(.032)
PROF_CHG	.016	−.005	.022	−.144	−.019
	(.011)	(.022)	(.037)	(.257)	(.016)
POS_SENTxPnL	1.435	−6.422	18.361**	−18.103	8.176
	(1.952)	(7.428)	(8.059)	(38.216)	(11.301)
POS_SENTxPROF_CHG	.189	.103	−.254	−1.997	−.218
	(.133)	(.381)	(.383)	(3.571)	(.192)
Constant	−1.500	−2.074	1.143	−4.372**	−4.743***
	(1.402)	(1.372)	(2.028)	(2.174)	(1.507)
Observations	2,308	2,277	2,362	2,405	2,407
R^2	.396	.397	.432	.396	.392
Adjusted R^2	.359	.359	.398	.361	.356

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. VIBTW weights were estimated using 10-K* filings from the past 14 years (i.e., (Y-14) to (Y-1)). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 26: Annual Augmented MZ-Regression With Additional Filing-Related Control Variables (Extending Training Set)

	Dependent variable: PFRV				
Y	2013	2014	2015	2016	2017
TS_FC	.362***	.251***	.206***	.235***	.148***
	(.021)	(.021)	(.021)	(.021)	(.021)
NEG_SENT	3.656	4.367	12.760***	7.516**	16.215***
	(2.780)	(5.547)	(3.620)	(3.768)	(4.057)
POS_SENT	1.607	1.070	1.693	5.050**	2.010*
	(1.335)	(1.299)	(1.228)	(2.213)	(1.175)
ASSERT	.215	.293	.655	−.671	−.573
	(.437)	(.473)	(.523)	(.643)	(.729)
UNCERT	.267	.727	−1.313	1.588	−.297
	(1.344)	(1.144)	(1.313)	(1.467)	(1.704)
LITI	7.827	6.015	−11.247	35.343**	37.596**
	(15.292)	(14.026)	(14.484)	(13.828)	(15.547)
GFS	−.020	.018	.022	.045*	−.011
	(.022)	(.025)	(.022)	(.024)	(.027)
FIN	.044*	.014	−.013	−.025	.032
	(.023)	(.027)	(.023)	(.025)	(.028)
SIZE	−.088***	−.129***	−.130***	−.126***	−.169***
	(.016)	(.015)	(.015)	(.015)	(.016)
BTM	.072***	.077***	.068***	.116***	.150***
	(.021)	(.020)	(.019)	(.019)	(.020)
TRVOL	.005	.026**	.041***	.071***	.053***
	(.011)	(.010)	(.011)	(.011)	(.012)
VIX	−.086***	.018	−.0004	.016	.045
	(.018)	(.014)	(.008)	(.011)	(.029)
LEVER	.223**	.173**	.332***	.430***	.475***
	(.089)	(.088)	(.081)	(.079)	(.085)
PnL_TURN	.263	.217	.212	.139***	−.153
	(.162)	(.244)	(.248)	(.050)	(.254)
SIZE_CHG	−.050	.086	.078	.010	.047
	(.066)	(.059)	(.056)	(.055)	(.052)
BTM_CHG	.019	−.065**	.034	.049	−.028
	(.035)	(.032)	(.032)	(.033)	(.034)
LEVER_CHG	−.016	−.004*	−.009	.004	−.013
	(.015)	(.003)	(.017)	(.016)	(.026)

SALES_GROWTH	−.001	.008	−.003	.001	.001
	(.001)	(.009)	(.002)	(.0005)	(.002)
SHR_BUYBACK	−.021	−.097***	−.084***	−.050	−.087***
	(.034)	(.034)	(.032)	(.031)	(.032)
PROF_CHG	.016	−.003	.006	−.003**	−.005
	(.011)	(.007)	(.014)	(.002)	(.004)
POS_SENTxPnL	1.435	.371	.521	−1.404	−2.045
	(1.952)	(1.940)	(1.769)	(2.858)	(1.641)
POS_SENTxPROF_CHG	.189	−.033	.059	.530**	−.038
	(.133)	(.063)	(.103)	(.233)	(.034)
Constant	−1.500	−1.972	.685	−5.712***	−5.156***
	(1.402)	(1.435)	(1.368)	(1.612)	(1.456)
Observations	2,308	2,277	2,362	2,405	2,407
R^2	.396	.397	.436	.404	.399
Adjusted R^2	.359	.359	.402	.369	.364

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. VIBTW weights were estimated using 10-K* filings from 1999 to (Y-1). Coefficients for boolean control variables (YRDUMMY, MTHDUMMY, WEEKDAYDUMMY, MONTHDAYDUMMY, SECTORDUMMY, and 10KDUMMY) are not displayed in the table.

Table 27: Augmented MZ-Regression With Additional Filing-Related Control Variables (Pooled OLS)

	Dependent variable: PFRV							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TS_FC	.248***	.251***	.252***	.252***	.252***	.251***	.251***	.253***
	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)
NEG_SENT	3.591***	.276***	.831	.395**	.575**	.189	.285	−.002
	(1.080)	(.096)	(2.254)	(.156)	(.228)	(.178)	(.267)	(.003)
POS_SENT	2.393***	.090	1.494	−.228	−.335	.140	.225	−.006
	(.593)	(.105)	(3.603)	(.256)	(.377)	(.162)	(.241)	(.009)
ASSERT	−.130	.430**	11.042	.475	.708	.464***	.678***	.074
	(.196)	(.186)	(7.776)	(.305)	(.452)	(.126)	(.187)	(.053)
UNCERT	.097	−.136	−.243	−.254	−.382	−.262*	−.411**	.002
	(.567)	(.182)	(1.295)	(.275)	(.403)	(.141)	(.207)	(.003)
LITI	15.277**	.071	.319	−.538**	−.695*	−.860***	−1.267***	.008
	(6.495)	(.067)	(17.362)	(.263)	(.382)	(.167)	(.246)	(.025)
GFS	.011	.006	.005	.006	.006	.003	.003	.007
	(.010)	(.010)	(.010)	(.010)	(.010)	(.010)	(.010)	(.011)
FIN	.008	−.001	.019	.015	.015	.018	.016	.029*
	(.011)	(.012)	(.013)	(.011)	(.011)	(.017)	(.017)	(.015)
SIZE	−.130***	−.146***	−.147***	−.146***	−.146***	−.150***	−.150***	−.148***
	(.007)	(.006)	(.006)	(.006)	(.006)	(.006)	(.006)	(.006)
BTM	.096***	.103***	.098***	.099***	.099***	.097***	.097***	.097***
	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)	(.009)
TRVOL	.037***	.041***	.043***	.042***	.042***	.041***	.041***	.042***
	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)	(.005)

VIX	.016***	.015***	.015***	.015***	.015***	.015***	.015***	.015***
	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
LEVER	.299***	.296***	.277***	.281***	.282***	.267***	.268***	.270***
	(.037)	(.037)	(.037)	(.037)	(.037)	(.037)	(.037)	(.037)
PnL_TURN	.120*	.129***	.181***	.146***	.144***	.152***	.151***	.181***
	(.069)	(.036)	(.017)	(.039)	(.039)	(.033)	(.035)	(.017)
SIZE_CHG	.037	.052**	.053**	.051**	.051**	.051**	.050**	.053**
	(.025)	(.025)	(.025)	(.025)	(.025)	(.025)	(.025)	(.025)
BTM_CHG	−.002	−.004	−.001	−.002	−.002	−.002	−.002	−.001
	(.014)	(.014)	(.014)	(.014)	(.014)	(.014)	(.014)	(.014)
LEVER_CHG	−.005**	−.005**	−.005**	−.005**	−.005**	−.005**	−.005**	−.005**
	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
SALES_GROWTH	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0002
	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)	(.0004)
SHR_BUYBACK	−.074***	−.079***	−.080***	−.080***	−.080***	−.078***	−.078***	−.081***
	(.015)	(.015)	(.015)	(.015)	(.015)	(.015)	(.015)	(.015)
PROF_CHG	.0003	−.001	.0003	.0003	.0002	.001	.001	.0003
	(.001)	(.001)	(.001)	(.002)	(.002)	(.001)	(.001)	(.0005)
POS_SENTxPnL	−.547	.266	9.439	.477	.721	.118	.168	.009
	(.838)	(.171)	(8.155)	(.411)	(.606)	(.117)	(.175)	(.007)
POS_SENTxPROF_CHG	.007	.001	−1.282	−.006	−.007	−.005	−.008	−.007
	(.019)	(.003)	(2.698)	(.024)	(.036)	(.007)	(.011)	(.007)
Constant	−2.532***	−2.106***	−1.984***	−1.945***	−1.952***	−1.881***	−1.865***	−2.043***
	(.664)	(.607)	(.609)	(.608)	(.608)	(.620)	(.620)	(.611)
Observations	11,759	11,759	11,759	11,759	11,759	11,759	11,759	11,759
R^2	.366	.365	.363	.364	.364	.365	.365	.363

Adjusted R^2	.358	.357	.356	.356	.356	.358	.358	.356

Notes: ***, **, and * denotes statistical significance at the one-, five- and ten-percent level, respectively. Standard errors are displayed in parentheses. Coefficients for boolean control variables (`YRDUMMY`, `MTHDUMMY`, `WEEKDAYDUMMY`, `MONTHDAYDUMMY`, `SECTORDUMMY`, and `10KDUMMY`) are not displayed in the table. Column headers (1) through (8) refer to the weighting schemes in the following order: `VIBTW`, `TFIDF`, `RFIDF`, `WFIDF_1PLOG`, `WFIDF_LOG1P`, `WF_1PLOG`, `WF_LOG1P`, `TFMAX`. Each column is a pooled OLS regression, estimated for the whole out-of-sample period from 2013 to 2017. Weights for the `VIBTW`, `TFIDF`, `RFIDF`, `WFIDF_1PLOG`, and `WFIDF_LOG1P` are estimated from a fixed training set (1999-2012).

Figures

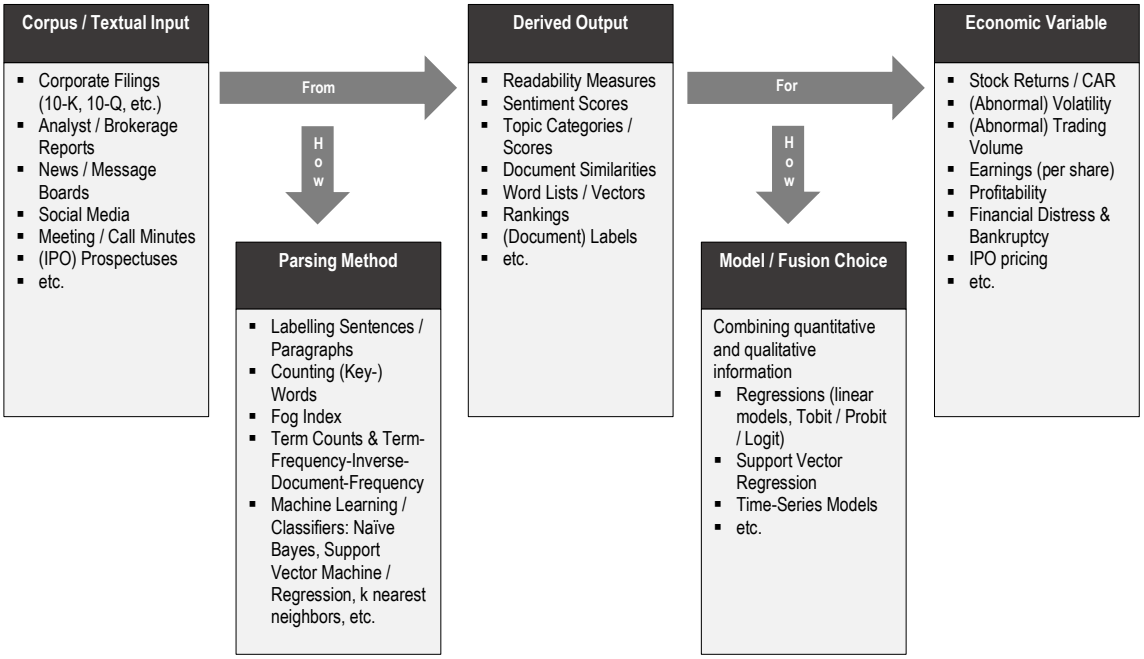


Figure 1: Schematic Representation of a Research Process: Textual Analysis in Finance

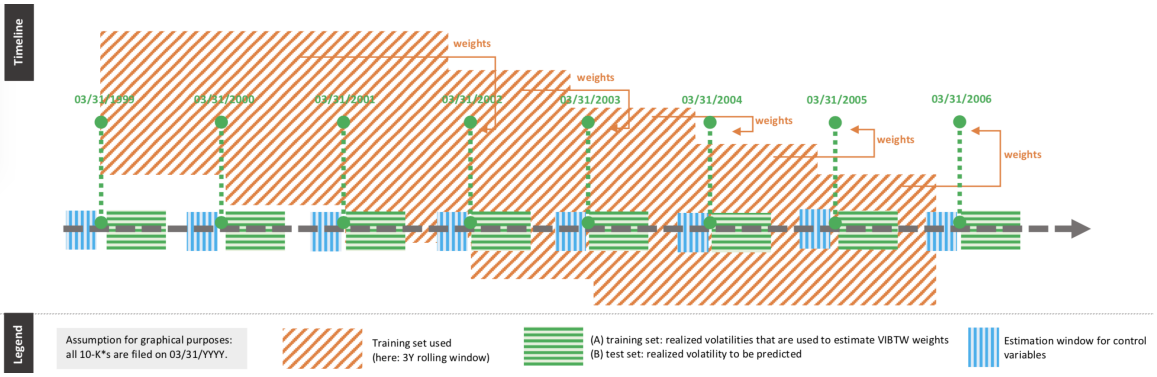


Figure 2: Visualization of the Research Design

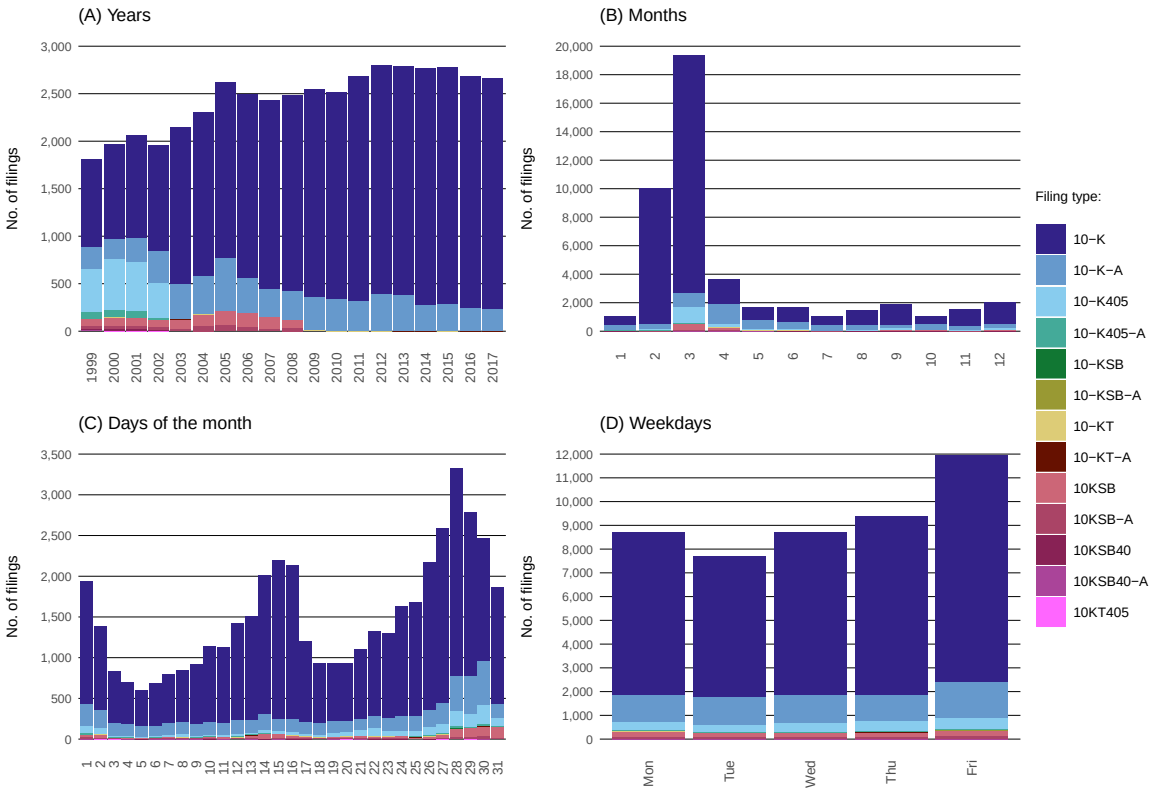
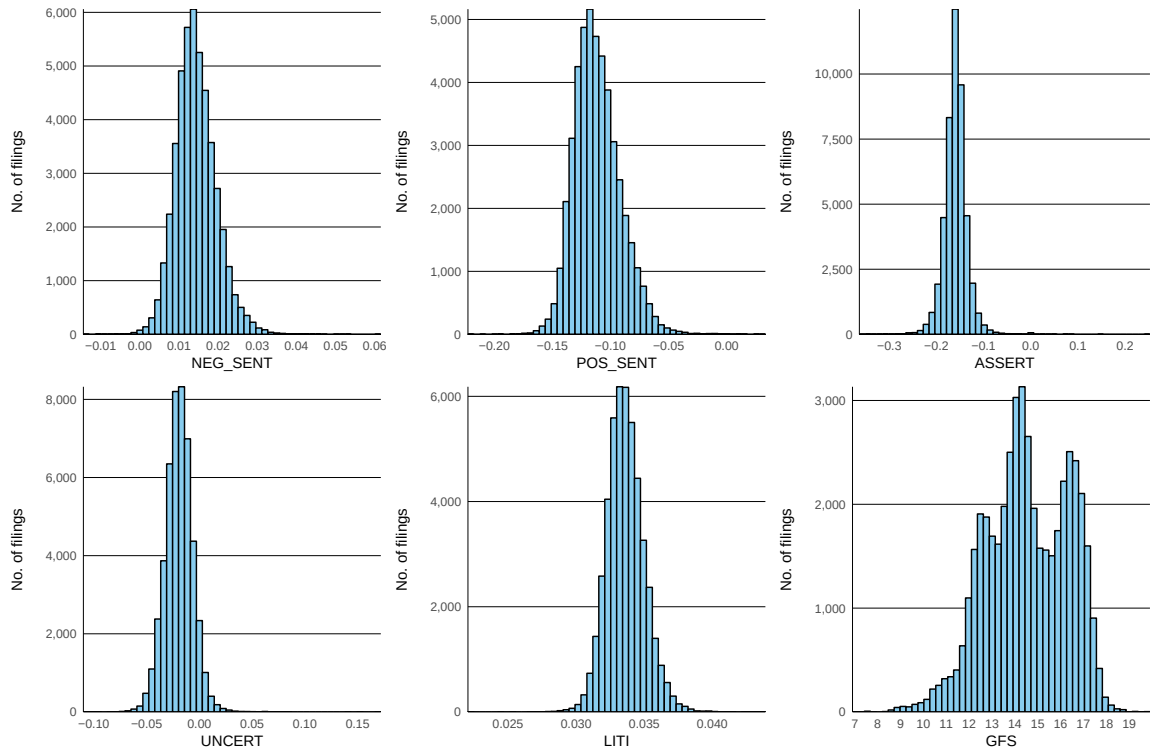
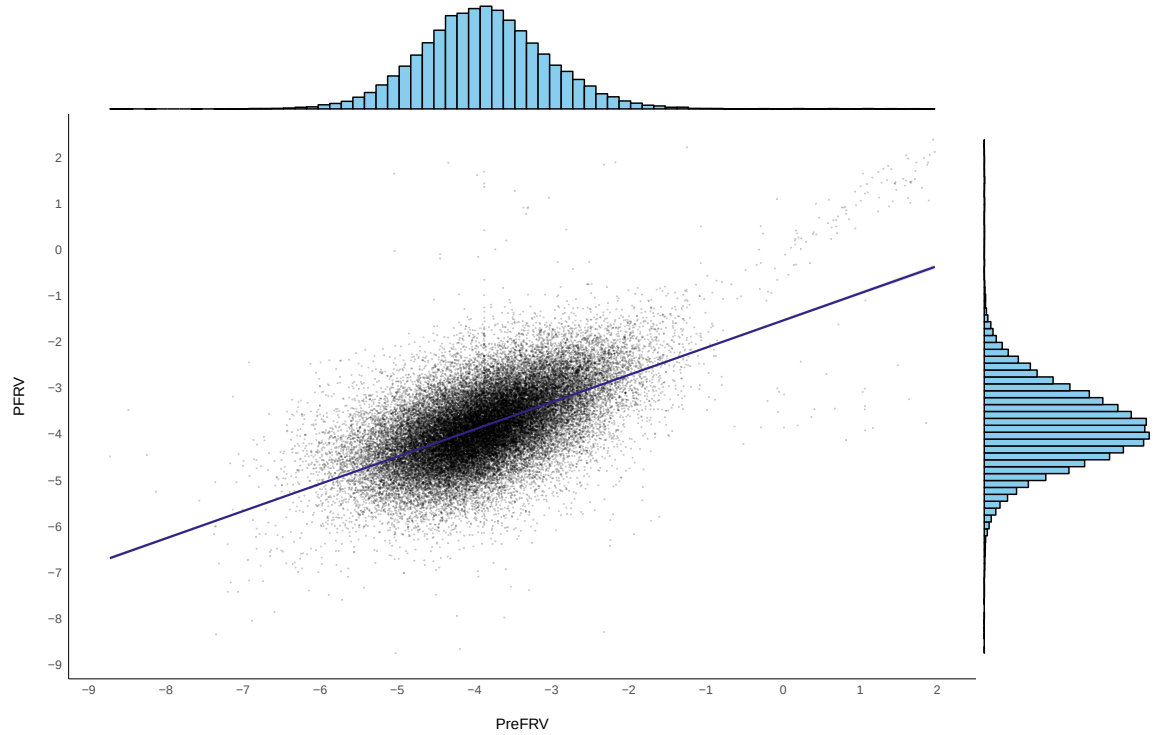


Figure 3: Sample Composition by Filing Year, Month, Day, and Type



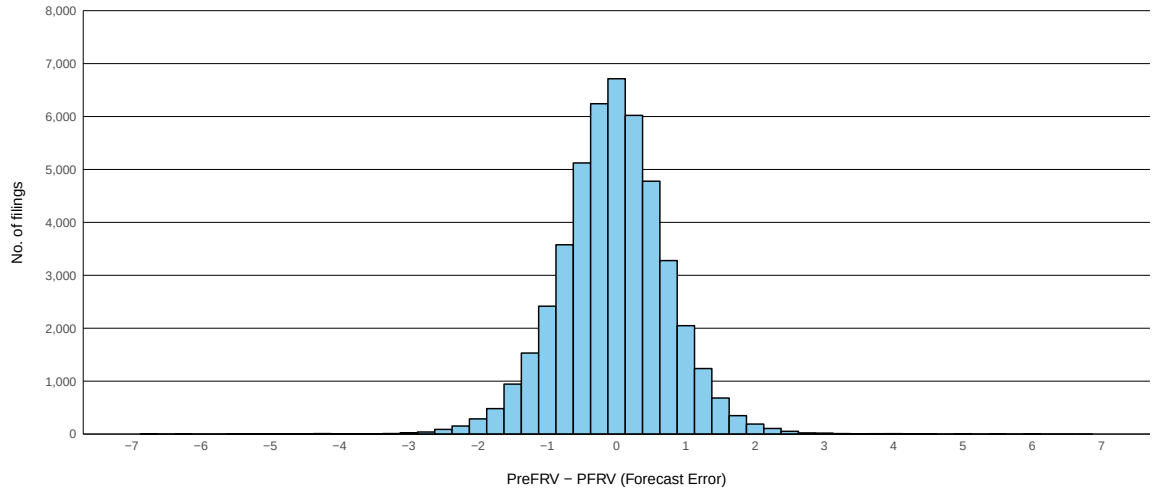
All variables (except for GFS) are VIBTW−scores with weights estimated from the full sample (i.e., 46,483 filings from 1999 to 2017).

Figure 4: Distribution of Text-Related Variables



The regression is estimated on variables in logarithmic form. Sample Size: 46,483.

Figure 5: Post- Versus Pre-Filing Realized Volatility : Scatterplot, Univariate MZ-Regression and Marginal Distributions



Differences are obtained from log−transformed volatility variables. Sample Size: 46,483.

Figure 6: Post- Versus Pre-Filing Realized Volatility: Forecast Error Distribution

A Appendix

A.1 LM Word Lists

The word lists are obtained from `https://sraf.nd.edu/textual-analysis/resources/` and are stemmed using the Porter Stemmer, implemented in the `tm` package in the statistical programming language `R`. The resulting lexica are composed as follows:

Negative ($J_N = 884$): abandon, abdic, aberr, abet, abnorm, abolish, abrog, abrupt, absenc, absente, abus, accid, accident, accus, acquiesc, acquit, acquitt, adulter, adversari, advers, aftermath, against, aggrav, alert, alien, alleg, annoy, annul, anomali, anomal, anticompetit, antitrust, argu, argument, arrearag, arrear, arrest, artifici, assault, assert, attrit, avers, backdat, bad, bail, bailout, balk, bankrupt, bankruptci, ban, bar, barrier, bottleneck, boycott, boycot, breach, *break*, breakag, breakdown, bribe, briberi, bridg, broken, burden, burdensom, burn, calam, calamit, cancel, careless, catastroph, caution, cautionari, ceas, censur, challeng, chargeoff, circumv, circumvent, claim, clawback, close, closeout, closur, coerc, coercion, coerciv, collaps, collis, collud, collus, complain, complaint, complic, compuls, conceal, conced, concern, concili, condemn, condon, confess, confin, confisc, conflict, confront, confus, conspiraci, conspir, conspiratori, contempt, contend, content, contenti, contest, contract, contradict, contradictori, contrari, controversi, convict, correct, corrupt, cost, counterclaim, counterfeit, countermeasur, crime, crimin, crise, crisi, critic, crucial, culpabl, cumbersom, curtail, cut, cutback, cyberattack, cyberbulli, cybercrim, cybercrimin, damag, dampen, danger, deadlock, deadweight, debar, deceas, deceit, deceiv, decept, declin, defac, defam, defamatori, default, defeat, defect, defend, defens, defer, defici, deficit, defraud, defunct, degrad, delay, deleteri, deliber, delinqu, delist, demis, demolish, demolit, demot, denial, deni, denigr, deplet, deprec, depress, depriv, derelict, derogatori, destabil, destroy, destruct, detain, detent, deter, deterior, deterr, detract, detriment, devalu, devast, deviat, devolv, difficult, difficulti, diminish, diminut, disadvantag, disaffili, disagre, disagr, disallow, disappear, disappoint, disapprov, disassoci, disast, disastr, disavow, disciplinari, disclaim, disclos, discontinu, discourag, discredit, discrep, disfavor, disgorg, disgrac, dishonest, dishonesti, dishonor, disincent, disinterest, disinterested, disloy, disloyalti, dismal, dismiss, disord, disparag, dispar, displac, dispos, dispossess, disproport, disproportion, disput, disqualif, disqualifi, disregard, disreput, disrupt, dissatisfact, dissatisfi, dissent, dissid, dissolut, distort, distract, distress, disturb, divers, divert, divest, divestitur, divorc, divulg, doubt, downgrad, downsiz, downtim, downturn, downward, drag, drastic, drawback, drop, drought, duress, dysfunct, eas, egregi, embargo, embarrass, embezzl, encroach, encumb, encumbr, endang, endanger, enjoin, erod, eros, errat, *er*, erron, error, err, escal, evad, evas, evict, exacerb, exagger, excess, exculp, exculpatori, exoner, exploit, expos, expropri, expuls, extenu, fail, failur, fallout, fals, falsif, falsifi, falsiti, fatal, fault, faulti, fear, feloni, fictiti, fine, fire, flaw, forbid, forbidden, forc, foreclos, foreclosur, forego, foregon, forestal, forfeit, forfeitur, forger, forgeri, fraud, fraudul, frivol, frustrat, fugit, gratuit, grievanc, grossli, groundless, guilti, halt, hamper, harass, hardship, harm, harsh, harsher, harshest, hazard, hinder, hindranc, hostil, hurt, idl, ignor, ill, illeg, illicit, illiquid, imbal, immatur, immor, impair, impass, imped, impedi, impend, imper, imperfect, imperil, impermiss, implic, imposs, impound, impractic, impract, imprison, improp, improprieti, imprud, inabl, inaccess, inaccuraci, inaccur, inact, inactiv, inadequaci, inadequ, inadvert, inadvis, inappropri, inattent, incap, incapacit, incapac, incarcer, incid, incompat, incompet, incomplet, inconclus, inconsist, inconveni, incorrect, indec, indefeas, indict, ineffect, ineffici, inelig, inequit, inequ, inevit, inexperi, inexperienc, inferior, inflict, infract, infring, inhibit, inim, injunct, injur, injuri, inordin, inquiri, insecur, insensit, insolv, instabl, insubordin, insuffici, insurrect, intent, interfer, interf, intermitt, interrupt, intimid, intrus, invalid, investig, involuntarili, involuntari, irreconcil, irrecover, irregular, irrepar, irrevers, jeopard, justifi, kickback, know, lack, lacklust, lag, laps, late, launder, layoff, lie, limit, linger, liquid, litig, lockout, lose, loss, lost, malfeas, malfunct, malic, malici, malpractic, manipul, markdown, misappl, misappli, misappropri, misbrand, miscalcul, mischaracter, mischief, misclassif, misclassifi, miscommun, misconduct, misdat, misdemeanor, misdirect, mishandl, misinform, misinterpret, misjudg, mislabel, mislead, misl, mismanag, mismatch, misplac, mispric, misrepres, misrepresent, miss, misstat, misstep, mistak, mistaken, mistrial, misunderstand, misunderstood, misus, monopolist, monopol, monopoli, moratoria, moratorium, mothbal, negat, neglect, neglig, nonattain, noncompetit, noncompli, nonconform, nondisclosur, nonfunct, nonpay, nonperform, nonproduc, nonproduct, nonrecover, nonrenew, nuisanc, nullif, nullifi, object, objection, obscen, obsolesc, obsolet, obstacl, obstruct, offenc, offend, omiss, omit, oner, opportunist, oppos, opposit, outag, outdat, outmod, overag, overbuild, overbuilt, overburden, overcapac, overcharg, overcom, overdu, overestim, overload, overlook, overpaid, overpay, overproduc, overproduct, overrun, overshadow, overst, overstat, oversuppli, overt, overturn, overvalu, panic, penal, penalti, peril, perjuri, perpetr, persist, pervas, petti, picket, plaintiff, plea, plead, pled, poor, pose, postpon, precipit, preclud, predatori, prejudic, prejud, prejudici, prematur, press, pretrial, prevent, problem, problemat, prolong, prone, prosecut, protest, protestor, protract, provok, punish, punit, purport, question, quit, racket, ration, reassess, reassign,

recal, recess, recessionari, reckless, redact, redefault, redress, refus, reject, relinquish, reluct, renegoti, renounc, repar, repossess, repudi, resign, restat, restructur, retali, retaliatori, retribut, revoc, revok, ridicul, riskier, riskiest, riski, sabotag, sacrific, sacrif, sacrifici, scandal, scrutin, scrutini, secreci, seiz, sentenc, serious, setback, sever, sharpli, shock, shortag, shortfal, shrinkag, shut, shutdown, slander, slippag, slow, slowdown, slower, slowest, slowli, sluggish, solvenc, spam, spammer, stagger, stagnant, stagnat, standstil, stolen, stoppag, stop, strain, stress, stringent, subject, subpoena, substandard, sue, *su*, suffer, summon, summons, suscept, suspect, suspend, suspens, suspicion, suspici, taint, tamper, tens, termin, testifi, threat, threaten, tighten, toler, tortuous, tragedi, tragic, traumat, troubl, turbul, turmoil, unabl, unaccept, unaccount, unannounc, unanticip, unapprov, unattract, unauthor, unavail, unavoid, unawar, uncollect, uncompetit, uncomplet, unconscion, uncontrol, uncorrect, uncov, undeliver, undeliv, undercapit, undercut, underestim, underfund, underinsur, undermin, underpaid, underpay, underperform, underproduc, underproduct, underreport, underst, understat, underutil, undesir, undetect, undetermin, undisclos, undocu, undu, unduli, uneconom, unemploy, uneth, unexcus, unexpect, unfair, unfavor, unfavour, unfeas, unfit, unforese, unforeseen, unforseen, unfortun, unfound, unfriend, unfulfil, unfund, uninsur, unintend, unintent, unjust, unjustifi, unknow, unlaw, unlicens, unliquid, unmarket, unmerchant, unmeritori, unnecessarili, unnecessari, unneed, unobtain, unoccupi, unpaid, unperform, unplan, unpopular, unpredict, unproduct, unprofit, unqualifi, unrealist, unreason, unrecept, unrecover, unrecov, unreimburs, unreli, unremedi, unreport, unresolv, unrest, unsaf, unsal, unsatisfactori, unsatisfi, unsavori, unschedul, unsel, unsold, unsound, unstabil, unstabl, unsubstanti, unsuccess, unsuit, unsur, unsuspect, unsustain, unten, untim, untrust, untruth, unus, unwant, unwarr, unwelcom, unwil, unwilling, upset, urgenc, urgent, usuri, usurp, vandal, verdict, veto, victim, violat, violenc, violent, vitiat, void, volatil, vulner, warn, wast, weak, weaken, weaker, weakest, will, worri, wors, worsen, worst, worthless, writedown, writeoff, wrong, wrongdo

Positive ($J_P = 145$): abl, abund, acclaim, accomplish, achiev, adequ, advanc, advantag, allianc, assur, attain, attract, beauti, benefici, benefit, best, better, bolster, boom, boost, breakthrough, brilliant, charit, collabor, compliment, complimentari, conclus, conduc, confid, construct, courteous, creativ, delight, depend, desir, despit, destin, dilig, distinct, dream, easier, easili, easi, effect, effici, empow, enabl, encourag, enhanc, enjoy, enthusiasm, enthusiast, excel, except, excit, exclus, exemplari, fantast, favor, favorit, friend, gain, good, great, greater, greatest, happiest, happili, happi, highest, honor, ideal, impress, improv, incred, influenti, inform, ingenu, innov, insight, inspir, integr, invent, inventor, leadership, lead, loyal, lucrat, meritori, opportun, optimist, outperform, perfect, pleasant, pleas, pleasur, plenti, popular, posit, preemin, premier, prestig, prestigi, proactiv, profici, profit, progress, prosper, rebound, recept, regain, resolv, revolution, reward, satisfact, satisfactorili, satisfactori, satisfi, smooth, solv, spectacular, stabil, stabl, strength, strengthen, strong, stronger, strongest, succeed, success, superior, surpass, transpar, tremend, unmatch, unparallel, unsurpass, upturn, valuabl, versatil, vibranc, vibrant, win, winner, worthi

Uncertainty ($J_U = 129$): abey, almost, alter, ambigu, anomali, anomal, anticip, appar, appear, approxim, arbitrarili, arbitrari, assum, assumpt, believ, cautious, clarif, conceiv, condit, confus, conting, could, crossroad, depend, destabil, deviat, differ, doubt, exposur, fluctuat, hidden, hing, imprecis, improb, incomplet, indefinit, indetermin, inexact, instabl, intang, likelihood, may, mayb, might, near, nonassess, occasion, ordinarili, pend, perhap, possibl, precaut, precautionari, predict, predictor, preliminarili, preliminari, presum, presumpt, probabilist, probabl, random, reassess, recalcul, reconsid, reexamin, reinterpret, revis, risk, riskier, riskiest, riski, rough, rumor, seem, seldom, sometim, somewhat, somewher, specul, sporad, sudden, suggest, suscept, tend, tentat, turbul, uncertain, uncertainti, unclear, unconfirm, undecid, undefin, undesign, undetect, undetermin, undocu, unexpect, unfamiliar, unforecast, unforseen, unguarante, unhedg, unidentifi, unknown, unobserv, unplan, unpredict, unprov, unproven, unquantifi, unreconcil, unseason, unsettl, unspecif, unspecifi, untest, unusu, unwritten, vagari, vagu, *vaguer*, *vaguest*, variabl, varianc, variant, variat, vari, volatil

Litigious ($J_L = 451$): abovement, abrog, absolv, access, acquire, acquiror, acquit, acquitt, addendum, adjourn, adjudg, adjud, adjudicatori, admiss, affidavit, affirm, affreight, aforedescrib, aforement, aforesaid, aforest, aggriev, alleg, amend, amendatori, anteced, anticorrupt, antitrust, anywis, appeal, appel, appelle, appointor, appurten, arbitr, arrearag, ascend, assert, assign, assum, attest, attorn, attorney, bail, baile, bailiff, bailment, benefici, bona, bonafid, breach, cedant, certiorari, cession, chattel, choat, claim, claimabl, claimant, claimhold, clawback, codefend, codicil, codif, codifi, collus, compensatori, complain, condemnor, confiscatori, consent, conservatorship, constitut, constru, contest, contract, contracthold, contractil, contractu, contraven, contravent, controvert, convenien, convey, convict, cotermin, counsel, *countersignor*, countersu, countersuit, court, courtroom, crime, crimin, crossclaim, deced, declar, decre, defalc, defeas, defect, defend, defer, deleg, delegat, delegate, delege, demur, demurr, depos, deposit, derog, design, desist, detain, devise, disaffili, disaffirm, disposit, dispossess, dispossessori, distraint, distribute, docket, done, duli, eject, encumb, encumbr, encumbranc, endorse, enforc, escheat, escrow, estoppel, evidenti, evidentiari, exceed, excis, exculp, exculpatori, executor, executori, *executric*, executrix, extracontractu, extracorpor, extrajudici, faci, facto, feloni, fide, *forbad*, forbear, forebear, forfeit, forthwith, *forwhich*, fugit, further, grantor, henceforth, henceforward, hereaft, herebi, heredita, herefor, herefrom,

herein, hereinabov, hereinaft, hereinbefor, hereinbelow, hereof, hereon, hereto, heretofor, hereund, hereunto, hereupon, herewith, herewithin, immateri, implead, inasmuch, incapac, incarcer, inchoat, incontest, indemnifi, indemnif, indemnite, indemn, indemnitor, indict, indorse, inforc, infract, infring, injunct, insofar, interlocutori, interplead, interpos, interposit, interrog, interrogatori, intestaci, intest, irrevoc, joinder, judici, judiciari, juri, jurisdict, jurisprud, jurist, juror, *juryman*, justic, law, lawmak, lawsuit, lawyer, legal, legales, legate, legisl, legislatur, libel, licens, lienhold, litig, litigi, *majeur*, mandamus, mediat, misdemeanor, misfeas, mistrial, moreov, motion, mutandi, nolo, nonappeal, nonbreach, nonconting, noncontract, noncontractu, noncontributori, nonfeas, nonfiduciari, nonforfeit, nonforfeitur, nonguarantor, noninfring, nonjudici, nonjurisdict, nonsever, nontermin, nonusuri, notari, notar, notwithstand, novo, nullif, nullifi, nulliti, oblige, obligor, offens, offere, offeror, optione, overrul, para, pari, passu, patente, pecuniarili, perjuri, permitte, perpetr, personam, petit, petition, plaintiff, plead, plea, pledge, pledgor, possessori, postclos, postclosur, postcontract, postjudg, preamend, predeceas, prehear, prejudic, prejud, prejudici, prepetit, presumpt, pretrial, prima, priviti, probat, probationari, *probation*, promulg, prorata, prorat, prosecut, prosecutor, prosecutori, proviso, punish, quitclaim, rata, ratabl, reargument, rebut, rebutt, record, recoup, recours, rectif, recus, redact, referenda, referendum, refil, regul, regulatori, rehear, reheard, release, remand, remedi, remis, repledg, replevin, request, requestor, reregul, rescind, resciss, restitutionari, retend, retroced, retrocessionair, revoc, rule, sentenc, sequestr, settlement, sever, shall, statut, statutorili, statutori, subclaus, subdocket, sublease, subleasehold, sublessor, sublicense, sublicensor, subparagraph, subpoena, subrog, subtrust, sue, *su*, summon, summons, supersed, supersedea, sureti, tenant, termin, terminus, testamentari, testifi, testimoni, thenc, thenceforth, *thenceforward*, thereaft, thereat, therefrom, therein, thereinaft, thereof, thereon, thereov, thereto, theretofor, thereund, thereunto, thereupon, therewith, tort, tortious, transferor, unappeal, unapp, unconstitut, uncontract, undefeas, undischarg, unencumb, unenforc, unlaw, unremedi, unstay, unto, usuri, usurp, vende, verdict, viatic, violat, voidabl, void, warrante, warrantor, whatev, whatsoev, whensoev, whereabout, wherea, whereat, wherebi, wherefor, wherein, whereof, whereon, whereto, whereund, whereupon, wherewith, whistleblow, whomev, whomsoev, whosoev, wil, will, wit, writ

Constraining ($J_C = 57$): abid, bound, commit, compel, compli, compuls, compulsori, confin, constrain, constraint, coven, depend, dictat, direct, earmark, encumb, encumbr, entail, entrench, escrow, *forbad*, forbid, forbidden, impair, impos, imposit, indebt, inhibit, insist, irrevoc, limit, mandat, mandatori, manditorili, necessit, noncancel, oblig, obligatori, permiss, permit, pledg, preclud, precondit, preset, prevent, prohibit, prohibitori, refrain, requir, restrain, restraint, restrict, stipul, strict, stricter, strictest, unavail

Strong Modal ($J_{SM} = 17$): alway, best, clear, definit, highest, lowest, must, never, strong, unambigu, uncompromis, undisput, undoubt, unequivoc, unparallel, unsurpass, will

Modest Modal ($J_{MM} = 13$): can, frequent, general, like, often, ought, probabl, rare, regular, should, tend, usual, would

Weak Modal ($J_{WM} = 18$): almost, appar, appear, conceiv, could, depend, may, mayb, might, near, occasion, perhap, possibl, seldom, sometim, somewhat, suggest, uncertain

This gives a total of 1,714 stemmed words; while due to overlapping in the lists, there are 1,574 *unique* words. The 140 words that appear in more than one list are underlined. 15 words (13 unique types) out of this list do not appear in any of the 46,483 10-K* documents of the corpus; they are printed in italic font. Details about the composition of LM lists and their respective size after stemming and searching for them in the corpus are provided in Table 28.

Table 28: Summary Statistics: LM Lexica

LM lexicon	k	Number of words		
		original	stemmed	occurring in corpus
Constraining	C	184	57	56
Litigious	L	903	451	442
Modest Modal	MM	14	13	13
Negative	N	2,355	884	882
Positive	P	354	145	145
Strong Modal	SM	19	17	17
Uncertain	U	297	129	127
Weak Modal	WM	27	18	18
Total		4,153	1,714	1,700
Uniques		3,887	1,574	1,562