Application of Textual Sentiment Scores in Value-at-Risk models

By

Jaime Casigay

Thesis submitted in partial fulfilment for the requirements of the degree of Master of Science Financial Engineering (M. Sc)

Department of Decision Sciences

HEC Montreal

December 2021

Abstract

We analyze the impact of textual sentiment as a predictor for market risk by constructing a sentiment score and applying it to Conditional Autoregressive Value-at-Risk (CAViaR) models of Engle & Manganelli (2004). Term-frequency data of words found in newswires, newspapers and web-publications concerning individual publicly-traded companies in the S&P 500 over seventeen years (1999-2016) is used to calibrate a sentiment scoring model via linear and sparse quantile regression (Koenker & Bassett Jr, 1978). Using the sentiment score as an external regressor for CAViaR models, Value-at-Risk (VaR) backtesting methods including the Dynamic Quantile test, Quantile Loss, and Actual-over-Exceedance ratio are used to evaluate model performance for one hundred companies with the highest frequency of publications over the time period. We conclude that there is a marginal improvement in predictive power over baseline models from our textual sentiment score for higher levels of VaR (1%).

Acknowledgements

I want to thank my professors and professional mentors who guided me throughout this rigorous academic process. I want to thank Keven Bluteau, Linda Mhalla and my supervisor, David Ardia, for their insighful input and incredible levels of patience. Undertaking this thesis has contributed enormously to my academic and personal growth.

I would also like to thank my friends and family for their support. They were always there to get me through the thick of things.

Table of Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 Introduction

A significant amount of attention has recently focused on the use of market sentiment to better understand financial markets. Significant periods of market volatility are often associated with sentiment such as market panic or excessive optimism. Value-at-Risk (VaR) models are broadly used by risk professionals and academics to provide a baseline assessment of the potential risk arising from their positions in financial markets. There is widespread interest in the academic and practitioner community for indicators that provide a measure of market sentiment and its potential to help predict market characteristics such as asset returns or volatility. Traditional measures of market sentiment have ranged from price and volume data, the measure of implicit volatility in options trading and surveys targeting consumer and business confidence. Outside of these traditional sources, Big Data, which encompasses textual, audio and visual information, has also become a substantial source of information. Big Data provides its own challenges including the high presence of noise and a lack of consistency or regularity in data that demands considerable data processing. One major source of Big Data sentiment analysis has focused on the use of text and its quantification to obtain valuable information that is not obtained directly from market price data; see Gentzkow *et al.* (2019) and Algaba *et al.* (2020). This thesis will apply econometric methods to construct a sentiment score indica-

tor based on textual data and apply it to conditional autoregressive Value-at-Risk (CAViaR) models; see Engle & Manganelli (2004) .

Sentiment is a latent variable that cannot be observed directly and hence, proxy variables are created in order to measure it. In the domain of natural language processing, sentiment analysis assesses the implied sentiment associated with specific words, sentences and overall bodies of text or spoken word. The use of sentiment towards econometric models requires the collection, processing and construction of aggregate quantitative measures that can be used as variable inputs. A body of text can be interpreted as a set of sentences that are themselves sets of words. The sets and their elements can each be individually represented as numerical vectors. These vectors can be aggregated in order to construct an overall matrix, a document matrix, representing information about the text; see Gentzkow *et al.* (2019).

Natural language processing is a vast domain of research ranging in its use of advanced statistical, econometric and machine learning methods towards the understanding of text and sentiment. Language is complex given the interactions of different words, negations, and phrases that contribute towards defining sentiment. This thesis centers its focus on the contribution of individual words to sentiment rather than complete phrases. Sentiment assesses an entity's expression of disposition towards others or even itself via a communication medium. Sentiment can also have an associated polarity such as positive versus negative messages, otherwise known as tone. Such a process can involve associating words as positive or negative to assessing the combined effects of those words in determining the sentiment for sentences and larger bodies of text; see Tetlock *et al.* (2008), Jegadeesh & Wu (2013), and Algaba *et al.* (2020).

Within finance, initial researchers (Das & Chen, 2007; Tetlock, 2007; Loughran & McDonald, 2011) would categorize words as positive or negative and assess the tone of a text through an adjusted frequency of positive versus negative words. They created and applied lexicons as signifiers of textual tone. A lexicon is a dictionary of language whereby sentiment can be directly labeled to each word or certain groupings

of words. More complex models accounted for measurement of the contribution of individual words to returns with Jegadeesh & Wu (2013) applying the concept of linear regression of adjusted word frequencies directly to stock returns.

Previous research (Garcia, 2013; Wisniewski & Lambe, 2013; Hanna *et al.,* 2020) have pointed towards the importance of textual sentiment in the understanding of short and sudden tail events such as bubbles and crashes. In behavioural economics, Baker & Wurgler (2007) and Shiller (2020) propose the hypothesis that the media can perpetuate narratives which can influence noise traders and lead to a self-fulfilling prophecy of market events. The importance of sentiment in extreme events outlines the need to apply methods which specifically aim to study tail risk.

This thesis contributes to the literature by first applying a novel method, quantile regression, to construct a sentiment score with inspiration taken from the work of Jegadeesh & Wu (2013), and Ardia, Bluteau, *et al.* (2019). Quantile regression proves useful in studying events at percentiles of distributions as opposed to ordinary linear regression that estimates the mean. Second, this thesis contributes to the literature by applying the score as a regressor for Conditional Autoregressive Value-at-Risk (CAViaR) models of stock returns. CAViaRs are applied individually on the returns of a hundred companies and it is shown that the score provides additional predictive information in assessing the quantile of returns versus only using pricing history. VaR backtesting methods including the dynamic quantile test (Engle & Manganelli, 2004) and quantile loss (Allen *et al.,* 2005) are used to assess the augmented CAViaR models against their returns-only CAViaR benchmarks.

1.2 Sentiment and Measurement

Keynes (1936) proposes the idea of *animal spirits* in financial markets which evolves into the field of behavioural finance by Shiller (2000). Market participants such as investors are influenced not only by firm fundamentals but also by their interactions and communications with each other and non-investment actors such as the media and company management. The communication of narratives can exacerbate

market sentiment and result in reactive crowd behaviour. Emotional words such as *exuberance* or *fear* are commonly used in financial communications whether by media, market analysts or company management to gauge and communicate the overall feeling or sentiment of market participants.

A definition of sentiment should be established in order to study its impact in different economic and social environments. There is general agreement that sentiment involves the communication of disposition (e.g. mood or inclination). Algaba *et al.* (2020) propose that sentiment is the disposition of an entity towards an entity, expressed via a certain medium. Liu *et al.* (2010) provides a similar definition defining an opinion holder that expresses an opinion, either direct or comparative towards an object. First, sentiment is an entity's expression of disposition via a communication medium. In a financial context, this can range from an investor's outlook being expressed via Twitter or newspapers, a textual medium, to direct opinions expressed via video platforms such as cable news outlets. This thesis limits itself to studying the sentiment from sources of text. Second, the disposition has a measurable polarity or semantic orientation. This can range from positive to negative in general, dovish versus hawkish with respect to central banks (Picault & Renault, 2017) or bullish versus bearish for financial markets (Antweiler & Frank, 2004). Third, the sentiment is oriented towards (an aspect of) another entity, or exceptionally the expressing entity itself. People communicate their ideas or sentiment to fellow investors, financial professionals, the general public, etc. Finally, sentiment can be associated within a given time frame as to when it was expressed.

Previous research has shown that textual sentiment has an impact on financial returns and volumes. The most dominant ideas within the literature are that textual sentiment has an impact in the immediate short-term (Chan, 2003; Tetlock, 2007; Da *et al.,* 2011; Jegadeesh & Wu, 2013), and it is most associated with negative returns or extreme events such as recessions (Garcia, 2013; Ahmad *et al.,* 2016; Hanna *et al.,* 2020).

One of the primary challenges in the applications of econometric methods to-

wards sentiment analysis is the derivation of a quantitative measure from qualitative data. Traditional quantitative measures in financial academia such as daily returns or trading volumes are unambiguous. This is a direct contrast to sentiment measurement whereby the choice of measure is dependent on the researcher. Language is multidimensional and complex considering elements such as the direct definition of words, their meanings under a specific context (e.g. financial or psychosocial) and the possible interactions between words (e.g. negations, superlatives) and sentences. Approaches to this complexity can involve the simplifying assumption of independence between words to the application of complex machine learning methods (e.g. neural networks) that assess the order and recurrence of words in sentences. Therefore, in the quantification of text, the choice of methodology and measure can facilitate a computational simplicity (e.g. normalized word frequency counts) at the expense of foregoing additional information and consequently, introducing noise to the measure (Loughran & McDonald, 2016).

The Bag-of-Words approach is a simple approach that assumes independence between words, meaning the order or syntax, and thus direct context, is unimportant. For Bag-of-Words, textual information is summarized directly as the counts of individual words. This permits high dimensional groups of words (e.g. sentences, paragraphs, documents) to be reduced to a term document matrix where the individual words can be organized into columns alongside their respective word counts as rows. A normalization of the term document matrix to the total number of words can then take place in order to compare one document to another.

Given the relative frequency of words in a document, various methods have been applied to derive the meaning or sentiment of them. One of the earliest and most used methods is the creation of lexicons or specific word lists. A lexicon can be considered as a dictionary for a specified purpose in research. Sentiment-based lexicons are lists of words that have been grouped to general degrees of sentiment such as positive or negative. This is often done via manual labeling of vocabulary. The

Harvard General Inquirer (GI) [1] word lists were developed for sociological and psychological research and have been used and adapted extensively in finance (Tetlock, 2007; Tetlock *et al.,* 2008; Kothari *et al.,* 2009; Heston & Sinha, 2017). The GI group words into more than 100 attributes including need, pleasure, pain, political, and interpersonal relations. By using a predefined convention to assess the sentiment of a set of words, researcher subjectivity is avoided and the availability of the lexicon permits replicability of studies. The tone of a document can be measured by quantifying the overall percentage of words in a document that belong to the specified groups and whether a majority or a net contribution of words are either positive or negative.

One problem arising from the use of general word lists is that language can be context-specific. Words that are generally negative in an everyday social sense including *tax, cost, capital, board, liability* and *depreciation* are not necessarily negative within a financial setting and can be more seen as being matter-of-fact. Given the importance of context in language, researchers (Henry, 2008; Loughran & McDonald, 2011) have built context-specific lexicons suited for research in finance and economics. Typically, the analysis of accounting literature including company 10-K's and earnings press releases would be used to assess frequent and meaningful words in finance. The general conclusion is that the context-specific lexicons have contributed to a better measurement of tone compared to general lexicons.

Another challenge within textual sentiment is assessing the polarity of words and the concept of term-weighting. Term-weighting refers to the measure of importance and overall additional information for a certain word. Words such as *slump, recession, underperform, crisis* are generally viewed as negative but again, depending on the context, certain words can be more negative than others. A popular weighting scheme is the term-frequency - inverse document frequency, *tf-idf*, which is the product of two terms, the count of the specific word in a given document times the inverse count that the word appears across a set of documents (Luhn, 1958; Jones,

[1] Available at http://www.wjh.harvard.edu/ inquirer/homecat.htm

1972; Manning & Schutze, 1999). Although there is no theoretical backing to this approach (Gentzkow *et al.,* 2019), the logic is that a word's importance is tied to its frequency within a document as well as its rarity amongst documents. Applications of *tf-idf* include Loughran & McDonald (2011) who assessed the polarity of a document by giving equal weighting to positive and negative words and then, take the dominant sentiment or take a net sum of the positive and negative frequencies.

This work is inspired by Jegadeesh & Wu (2013) 's approach to term weighting and Ardia, Bluteau, *et al.* (2019) 's subsequent extension of their methodology. Adapting from the initial *tf-idf* methodology, Jegadeesh & Wu (2013) and Wu developed a term-weighting approach that assigns weights for each word based on market reactions to documents containing those words. They test multiple lexicons (Loughran & McDonald, 2011; Harvard IV-4 Psychosocial Dictionary; Bradley & Lang, 1999) and arrived at the conclusion that the specific lexicon is not as important as the overall term-weighting when assessing word importance.

In order to assign an aggregate sentiment score to a document, the relative term frequencies of words are regressed against the firm's returns. Their approach has the following intuitive properties:

1. The score is positively related to the number of occurrences of each positive and negative word.

2. The score is positively related to the strength of the negative or positive words.

3. The score is inversely related to the total number of words in the document.

For document i, Jegadeesh and Wu's score is the following:

$$Score_i = \sum_{j=1}^{J} (w_j F_{i,j}) \frac{1}{a_i} \tag{1.1}$$

where:

- w_j is the weight for word j that is estimated from linear regression
- $F_{i,j}$ is the number of occurences of word j in document i

- a_i is the total number of words in the document. The term $\frac{1}{a_i}$ reflects that the score is inversely related.

The weights are estimated indirectly through a regression against the abnormal stock return, (r_i), corresponding to company i. The abnormal stock return is the difference between the individual stock returns against the CRSP value-weighted index over three days.

$$\begin{aligned} r_i &= a + b\sum_{j=1}^{J}(w_j F_{i,j}\frac{1}{a_i}) + \epsilon_i \\ &= a + \sum_{j=1}^{J}(bw_j F_{i,j}\frac{1}{a_i}) + \epsilon_i \\ &= a + \sum_{j=1}^{J}(B_j F_{i,j}\frac{1}{a_i}) + \epsilon_i \end{aligned} \tag{1.2}$$

where a_i and $F_{i,j}$ can be computed directly and B_j is the regression coefficient which provides unbiased estimates of bw_j. Jegadeesh & Wu (2013) performed a subsequent regression using standardized estimates for the weights. The weight, $\hat{w_j}$, is equal to the difference between the slope coefficient estimate obtained from equation 1.3 and $\bar{B_j}$ across all words over the standard deviation of the estimated slope coefficient across all words.

$$\hat{w_j} = \frac{\hat{B_j} - \bar{B}}{StandardDeviation(\hat{B_j})} \tag{1.3}$$

Fitting the regression below using the standardized weights, they empirically arrive at $b > 0$.

$$r_i = a + b(\sum_{j=1}^{J}(\hat{w_j} F_{i,j})\frac{1}{a_i}) + \epsilon_i \tag{1.4}$$

Thus, Jegadeesh & Wu (2013) concluded that their tone measure conveyed incremental information to the market.

Ardia, Bluteau, *et al.* (2019) expand on Jegadeesh & Wu (2013)'s methodology by assessing the tone surrounding a particular firm, k. They generalize Jegadeesh and

Wu's score by studying the ensemble of documents relating to a specific firm versus assessing the tone from a single document for a collection of firms, $k = 1, ..., K$.

$$TONE_{k,t} = \sum_{j=1}^{J} \eta_j f(j,k,t) \tag{1.5}$$

$$f(j,k,t) = \frac{1}{D_{k,t}} \sum_{d=1}^{D_{k,t}} FQ_{d,j,k,t} \frac{1}{N_{d,k,t}} \tag{1.6}$$

where:

- J is the total number of words in the dictionary, $j = 1, ..., J$.
- η_j is the associated polarity score for word j.
- $f(j,k,t)$ is the average term frequency which maps a word j for firm k at time t to a real number.
- $D_{k,t}$ is the number of articles written about firm k at time t.
- $FQ_{d,j,k,t}$ is the number of times that the jth word is encountered in article d.
- $N_{d,k,t}$ is the number of words in article d.

The value for $D_{k,t}$ in Jegadeesh and Wu is 1 as they assessed for the tone of individual articles. Ardia, Bluteau and Boudt normalize the word frequency based on the total number of articles for firm k.

Both Jegadeesh & Wu (2013) and Ardia, Bluteau, *et al.* (2019) use linear regression in their approach to assess the weight/polarity score for the words in their respective dictionaries. This thesis expands on the methodology by applying quantile regression to study the weights of words at various quantiles. Quantile regression is useful because it does not rely on the distributional assumptions in linear regression. Second, the assessment of tone from the words is based on the specified quantile in the quantile regression. Indices taken from quantile regressions using the 1% and 5% quantiles are used as external regressors to CAViaR models to assess their informational value within a risk-management context.

1.3 Sentiment and Risk Management

A major aim of this thesis is to study the effect of textual sentiment on firm-level risk. Specifically, this thesis applies a proposed measure of sentiment to a standard risk model, CAViaR (Engle & Manganelli, 2004), to evaluate if the information from text adds value to predicting the Value-at-Risk (VaR) of firm-level returns. The use of quantile regression is a novel contribution to the literature with respect to sentiment analysis.

Given a response variable, Y, and a n-dimensional predictor, $\mathbf{x}$, with conditional cumulative distribution function $F_Y(y|\mathbf{x}) = \mathbb{P}(Y \leq y|x)$, the q-conditional quantile is defined as:

$$Q_q(Y|\mathbf{x}) = inf\{y : F_Y(y|\mathbf{x}) \leq q\} \tag{1.7}$$

Quantile regression is an extension of linear regression whereby the conditional q-quantile for a random response variable, Y, is predicted using a linear function across values of predictor variables, $\mathbf{x}$.

$$Q_q(Y|\mathbf{x}) \equiv \mathbf{x}'\beta(q), \quad 0 \leq q \leq 1 \tag{1.8}$$

$\beta(q)$ is a n-dimensional vector of regression coefficients dependent on the quantile level, q. Intuitively, given a sample of N observations $\{x_n, y_n\}$ for $n \, \epsilon \, \{1, 2, \ldots, N-1, N\}$, the quantile regression for a given quantile level, q, is a line such that the proportion of observations at or above the line is equal to $1 - q$ and the proportion of observations below the line is equal to q.

For a given quantile level, q, and linear model, $Q_q(Y|\mathbf{x}) \equiv \mathbf{x}'\beta(q)$, the parameter vector, $\beta(q)$ is estimated by minimizing the quantile loss function. The quantile loss function is the following:

$$\rho(\beta_q) = \sum_{i:y_i \geq x_i\beta_q} q|y_i - x_i\beta_q| + \sum_{i:y_i < x_i\beta_q} (1-q)|y_i - x_i\beta_q| \tag{1.9}$$

$$\rho(\beta_q) = \sum_{i=1}^{N} [\mathbf{1}_{\{y_i \geq x_i\beta_q\}} q(y_i - x_i\beta_q) + \mathbf{1}_{\{y_i < x_i\beta_q\}}(q-1)(y_i - x_i\beta_q)] \tag{1.10}$$

$$\rho(\beta_q) = \sum_{i=1}^{N} (q - \mathbf{1}_{\{y_i < x_i\beta_q\}})(y_i - x_i\beta_q) \tag{1.11}$$

This function is also known as the check function, $\rho(u)$ where $u = y_i - x_i\beta_q$.

$$\rho(u) \equiv \sum_{i=1}^{N} (q - \mathbf{1}_{\{u<0\}})u \tag{1.12}$$

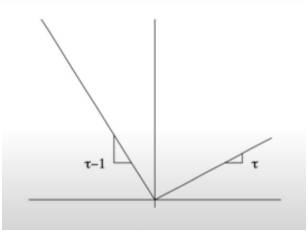

Figure 1.1: Quantile loss function

Claim: For a given quantile q, $\beta_q = \arg\min_\beta \mathbb{E}[\rho(y_i - x_i\beta_q)|x_i]$.

Consider the data-generating process,

$$\begin{aligned} y_t &= f(y_{t-1}, x_{t-1}, \ldots, y_1, x_1; \beta^0) + \epsilon_{tq} \qquad Quant_q(\epsilon_{tq}|\Omega t) = 0 \\ &\equiv f_t(\beta^0) + \epsilon_{tq} \end{aligned} \tag{1.13}$$

where $f_1(\beta^0)$ is some given initial condition.

Let $f_t(\beta) \equiv x_t\beta$. The q^{th} regression quantile is defined as any $\hat{\beta}$ that solves:

$$\min_\beta \frac{1}{T} \sum_{i=1}^{T} [q - \mathbf{1}_{\{y_t < f_t(\beta)\}}][y_t - f_t(\beta)] \tag{1.14}$$

Value-at-Risk is a quantile measure that reports the limit of losses associated with a specified probability and time period. For example, a 1-day 95%-quantile VaR of 1M\$ would indicate that within 95% of 1-day periods, losses should not

exceed 1M$. For the remaining 5% of 1-day periods, losses would exceed 1M$. Value-at Risk is dependent on assumptions of the return distribution (Damodaran, 2007). In practice, Value-at-Risk can be computed via:

- parametric or variance-covariance methods (e.g. GARCH)
- nonparametric (e.g. Historical simulation)
- semi-parametric (e.g. quantile regression, CAViaR, extreme value theory)

Value-at-Risk is closely linked to volatility models. Under the parametric approach, the estimated standard deviation of the return distribution from a volatility model is translated into the VaR quantile.
The conditional VaR for a series of returns, r_t, $t \epsilon \{0, 1, ..., T\}$, for a given quantile, q, is defined as follows:

$$\mathbb{P}[r_t < VaR_{t|t-1}(q)] = q, \quad \forall\, t \epsilon \mathbb{Z}, \quad q \epsilon (0, 1) \tag{1.15}$$

The economic literature supports the finding that textual sentiment and news releases has a measurable impact on stock price movements and volatility. Tetlock (2007) construct textual sentiment measures and find that the conditional volatility of the Dow Jones is higher when their constructed pessimism factor is high. Kothari *et al.* (2009) show asymmetric market responses to firm management's release of good and bad news and subsequently, how managers tend to delay the publication of bad news. Boudoukh *et al.* (2013) show that stock-level volatility is similar on no-news days and unidentified news days while on identified news days, the volatility of stock prices is over double that of other days. Banerjee *et al.* (2021) use sentiment word lists to construct measures and find that news sentiment is highly correlated to bond return volatility.

Sentiment measures derived from text can be applied as external regressors to improve various volatility models. Examples of models commonly cited in the literature and their use of external regressors include the Generalized autoregressive conditional heteroskedasticity model ($GARCH$) (Antweiler & Frank, 2004) and the

heterogenous autoregressive model (*HAR*) (Caporin & Poli, 2017; Audrino *et al.,* 2020; Lehrer *et al.,* 2021) alongside their various extensions. Antweiler & Frank (2004) applied a Naive Bayes algorithm to word counts on internet message boards and used their sentiment indicator to improve GARCH, EGARCH and GJR models. Caporin & Poli (2017), Audrino *et al.* (2020) and Lehrer *et al.* (2021) test large groups of market sentiment measures (including text-based ones) and optimally selected amongst them using a least absolute shrinkage and selection operator (*LASSO*) approach to improve extensions of the HAR model.

The use of CAViaR models alongside sentiment analysis is novel with no prior known research on its application. CAViaR models were proposed by Engle & Manganelli (2004) to directly model the quantile of returns. CAViaR models are autoregressive models whose parameters are optimized via quantile regression.

A generic CAViaR specification can be seen as an autoregressive function of its lagged quantile estimates and lagged external regressors:

$$f_t(\boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{q} \beta_i f_{t-i}(\boldsymbol{\beta}) + \sum_{j=1}^{r} \beta_j l_{t-j}(\mathbf{x}_{t-j}) \tag{1.16}$$

where:

- $f_t(\boldsymbol{\beta})$ is the quantile estimate of returns for a given quantile, q
- $l_t(x_t)$ is a series of additional regressor variables (e.g. lagged returns)

Engle & Manganelli (2004) propose four different types of CAViaR models. Each of these models used lagged returns as an additional regressor to the lagged quantile estimates.

- Symmetric absolute value (SAV) which responds symmetrically to lagged returns.

$$f_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 f_{t-1}(\boldsymbol{\beta}) + \beta_3 |r_{t-1}| \tag{1.17}$$

- Asymmetric slope (AS) with different responses to positive versus negative

lagged returns.

$$f_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 f_{t-1}(\boldsymbol{\beta}) + \beta_3 (r_{t-1})^+ + \beta_4 (r_{t-1})^- \tag{1.18}$$

- Indirect GARCH ($GARCH$) which responds symmetrically to lagged returns.

$$f_t(\boldsymbol{\beta}) = (\beta_1 + \beta_2 f_{t-1}^2(\boldsymbol{\beta}) + \beta_3 r_{t-1}^2)^{\frac{1}{2}} \tag{1.19}$$

- Adaptive which responds to past hits of VaR by increasing or decreasing the lagged quantile using positive hyperparameter G.

$$f_t(\boldsymbol{\beta}) = f_{t-1}(\boldsymbol{\beta}) + \beta_1 \{[1 + exp(G[r_{t-1} - f_{t-1}(\beta)])^{-1} - q]\} \tag{1.20}$$

Jeon & Taylor (2013) extend the original models proposed by Engle & Manganelli (2004) to include the *implied quantile* which is derived from options implied volatility as an additional regressor. We aim to replicate this approach but use a textual sentiment score measure as a regressor instead of the implied quantile.
To calibrate the CAViaR model, the parameter vector, $\hat{\boldsymbol{\beta}}$, is the argument that minimizes the quantile loss function in equation 1.14.

This thesis applies a constructed sentiment score measure to the symmetric absolute value (SAV), asymmetric slope (AS) and indirect GARCH(1,1) CAViaR models and applies VaR backtesting methods to compare performance between baseline and augmented models.

1.4 Value-at-Risk and Backtesting

The Value-at-Risk measure derived from any model is evaluated with respect to the hit indicator, $Hit_t(q)$, for a given quantile, q. $Hit_t(q)$ is a binary variable with the *ex-post* observation of a q% VaR violation at time t.

$$Hit_t(q) = \begin{cases} 1 & \text{if} \quad r_t < VaR_{t|t-1}(q) \\ 0 & \text{otherwise} \end{cases} \tag{1.21}$$

Christoffersen (1998) notes that VaR forecasts are valid if and only if they satisfy two hypotheses:

- Unconditional coverage (UC) hypothesis (Kupiec, 1995): The probability of an *ex-post* return exceeding the VaR forecast must be equal to the q coverage rate. Essentially, a VaR model that predicts the q-th quantile of returns should not over- or underestimate the quantile level of risk.

$$\mathbb{P}[Hit_t(q) = 1] = \mathbb{E}[Hit_t(q)] = q \tag{1.22}$$

- Independence (IND) hypothesis: In order to accurately model the higher-order dynamic of returns, the hit indicator, $Hit_t(q)$, at time t for violation rate q % should be independent of $Hit_{t-k}(q), \forall k \neq 0$. Past VaR violations should not be informative of present and future violations. A model that does not demonstrate the independence hypothesis can lead to clustering of VaR violations even if it has the correct average number of violations (Dumitrescu *et al.,* 2012).

If both the UC and IND hypotheses are satisfied, the VaR violation process is a martingale difference sequence and has correct conditional coverage (CC) under the information known at the previous time period, $\mathcal{F}_{t-1}$.

$$\mathbb{E}[Hit_t(q)|\mathcal{F}_{t-1}] = q \tag{1.23}$$

We test three major metrics to assess the performance of the CAViaR VaR forecasts:

- Dynamic Quantile (DQ) test (Engle & Manganelli, 2004)

- Quantile Loss Ratio (Koenker & Bassett Jr, 1978; González-Rivera *et al.,* 2004)

- Actual over Exceedance (AE) Ratio (Ardia, Boudt, *et al.,* 2018)

The DQ test uses a linear regression model to test the independence of the hit indicator, $Hit_t(q)$. Under CC, the conditional expectation of $Hit_t(q)$ given past information must be zero. It evaluates the following regression model:

$$\begin{aligned} Hit_t(q) =& \delta + \sum_{k=1}^{K} \beta_k Hit_{t-k}(q) \\ &+ \sum_{k=1}^{K} \gamma_k g[Hit_{t-k}(q), Hit_{t-k-1}(q), \ldots, Hit_1(q); \mathcal{F}_{t-1}] + \epsilon_t \end{aligned} \tag{1.24}$$

where ϵ_t is a discrete i.i.d process and $g(.)$ is a function of past hit indicators and the information set, $\mathcal{F}_{t-1}$. Testing for the joint nullity of the coefficients would therefore check for correct conditional coverage.

$$H_0 : \delta = \beta_1 = \cdots = \beta_k = \gamma_1 = \cdots = \gamma_k = 0, \quad \forall k \ = 1, \ldots, K \tag{1.25}$$

The quantile loss is the same as that used by Koenker & Bassett Jr (1978) for quantile regression. Specifically, for period t at quantile q, the quantile loss, $QL_t(q)$, is defined as:

$$QL_t(q) \equiv (q - 1_{\{r_t < f_t(\beta)\}})(r_t - f_t(\beta)) \tag{1.26}$$

with Equation 1.14 being the average quantile loss. Quantile loss is an asymmetric loss function that penalizes more heavily with weight (1 - q) the observations of VaR exceedance. Given two hit indicator series, A and B, the quantile loss ratio is the ratio between the average quantile losses for both A and B. If $QL_A/QL_B < 1$, then A outperforms B and vice versa (Ardia, Boudt, *et al.,* 2019).

The AE ratio tests for unconditional coverage. The AE ratio is defined as:

$$\frac{\sum_{t=1}^{T} Hit_t(q)}{\mathbb{E}[Hit_t(q)]T} = \frac{\sum_{t=1}^{T} Hit_t(q)}{qT} \quad \text{under UC} \tag{1.27}$$

The closer the absolute value of the AE ratio is to 1, the better the model. An AE ratio < 1 is considered too conservative with the model making less hits than expected and an AE ratio > 1 is considered to underestimate the risk with the model making more hits than expected.

Overall, we aim to look for improvements between baseline models and models that are augmented with our constructed sentiment score for the three VaR back-testing metrics.

Chapter 2

Data

The data composes of three major components:

1. Individual daily stock returns for 598 companies. Out of 598 companies, only 100 companies are selected for VaR analysis based on those with the most average term frequency observations (equation 1.6).

2. Fama and French research factors for the Fama-French five factor model.

3. Average term frequencies for individual firms based on news articles about firm k at time t.

2.1 Returns

Across 598 firms, the news sample spans from January 1, 1999 to December 31, 2016. Individual daily stock returns are sourced from the CRSP/Compustat database.

2.2 Fama and French Factors

Fama and French factors are sourced from Kenneth French's Dartmouth college website (French, 2013). Specifically, the five-factor model (Fama & French, 2015) was used. Fama and French factors (Fama & French, 1993) are time series of long-short portfolio strategy returns based on company fundamentals. They are used as

control variables representing systematic market factors in deriving our sentiment indicator (section 3.1). For a given time t, the associated vector of factors:

$$FF_t = [MKT_t, SMB_t, HML_t, RF_t, UMD_t]' \tag{2.1}$$

with MKT_t being the market risk premium, SMB_t being the small-minus-big factor, HML_t being the high-minus-low factor, RF_t being the risk-free rate and UMD_t being the momentum factor all at time t.

2.3 Average Term Frequency

The average term frequency (equation 1.6) was obtained from a constructed database obtained from Ardia, Bluteau, *et al.* (2019). They performed a prior analysis on news articles retrieved from LexisNexis discussing 598 non-financial firms that were included in the S&P 500. Filters and controls were done for relevance score, the type of publication (i.e. newswire, newspaper, web publication), the presence of duplicates, individual company focus instead of overall industry focus and human versus machine-written texts. News is also controlled such that it is media-sourced and not sourced from the actual firm in order to account specifically for media sentiment.

Their sample contains over 2,315,402 news articles, with an average of 3,871 articles per firm and an average daily coverage of 34.81% meaning that approximately at least one article is published for a given firm once every three days. The database covers 3,585 words. This lexicon was obtained by merging the Loughran & McDonald (2011) and Harvard IV-4 (Stone & Hunt, 1963) lexicons.

Chapter 3

Methodology

3.1 Sentiment Score Construction

For each individual company, a sentiment score is constructed via a two-step regression; first step: linear regression, second step: quantile regression. The series of individual returns are first regressed against their corresponding Fama-French factors and then against the average term frequencies. This method was chosen as opposed to one-step quantile regression due to repeating values for the Fama-French factors at the same date for different firms and hence, if a combined Fama-French and term frequency panel was used, the result would be a singular matrix for the quantile regression loss function . The objective is to obtain a sentiment score that accounts only for idiosyncratic differences in the media tone and coverage.

The two regressions and subsequent sentiment score are outlined below.

3.1.1 Regression 1: Linear Regression

The series of returns of company i ($i \in \{1, 2, ..., k\}$) is regressed to the corresponding Fama-French factors using linear regression.

$$R_i = \alpha_i + \beta_i' FF_i + \eta_i \tag{3.1}$$

where:

- α_i is a constant for company i.
- R_i is a $T_i \times 1$ vector of daily stock returns for company i with a time frame of length T_i.
- FF_i is a $T_i \times 5$ matrix of the corresponding five Fama-French factors over the time frame for company i.
- β_i is a 5×1 vector of coefficients to the five Fama-French factors as outlined in equation 2.1.
- η_i is a $T_i \times 1$ vector of residuals which will be the regressand for the second quantile regression at quantile q (section 3.1.2).

Therefore, the purpose of β_i is to account for the coefficient of company i to systematic market factors. The vector of residuals, η_i, would therefore be linked to idiosyncratic factors affecting company i. The aim of the second regression is to assess whether additional predictive power occurs from the average word frequencies.

3.1.2 Regression 2: Quantile Regression

For k companies, $i \epsilon \{1, 2, ..., k\}$, the residuals, η_i, from regression in section 3.1.1 were concatenated and regressed to the corresponding average term frequency via quantile regression using the R quantreg package and the sparse quantile regression function, rq.fit.sfn (Koenker, Portnoy, *et al.,* 2018).

The objective is to find a set of vocabulary coefficients, Λ_q, to estimate the quantile of residuals. We take the quantile estimate of residuals as our sentiment score. Considering the residuals from equation 3.1 contain idiosyncratic information about the returns of company i, modeling the quantile estimate would represent modeling deviations from the market. To calibrate Λ_q, we define the linear function as follows and use quantile regression to find the argument which minimizes the quantile loss function (1.14) for:

$$f_t(\Lambda_q) \equiv \begin{bmatrix} Freq_{1,t-1} \\ Freq_{2,t-1} \\ \vdots \\ Freq_{k-1,t-1} \\ Freq_{k,t-1} \end{bmatrix} \Lambda_q \tag{3.2}$$

where:

- $f_t(\Lambda_q)$ is the quantile estimate of the residuals, η_i at time t
- $Freq_{i,t-1}$ is a $T_i \times V$ matrix of *one-day lagged* $(t-1)$ average term frequencies (1.6) for a lexicon vocabulary of V words across time frame T_i for company i
- Λ_q is a $V \times 1$ vector of vocabulary coefficients, $\lambda_{v,q}$, to the average term frequencies for the specified quantile, q.

The sentiment score is defined as:

$$SCORE_{q,i,t} = \sum_{v}^{V} \lambda_{q,v} f(v, i, t-1) \tag{3.3}$$

where:

- $SCORE_{q,i,t}$ is a measure of sentiment for company i at time t fitted to quantile q of returns.
- $\lambda_{q,v}$ is the coefficient for the average term frequency of word v amongst a vocabulary of V words fitted to quantile q of returns.
- $f(v, i, t-1)$ represents the average term frequency (1.6) of word v amongst media publications about company i at time $t-1$.

This one-day lagged quantile regression aims to test whether the average term frequencies possess predictive power for a given quantile of residuals and indirectly, the returns. $SCORE_{q,i,t}$ is estimated via quantile regression using the sparse implementation of the Frisch-Newton interior point algorithm described in Portnoy & Koenker (1997).

The data was divided using an expanding window to calibrate the model. With the exception of the first year of data, for a given year, the sentiment coefficients were calibrated using data up to one year before. For example, the sentiment score constructed in 2009 would use data from 1999-2008 as its window to obtain estimated sentiment coefficients. Hence, from 1999 - 2016, there would be 17 expanding windows and 17 estimates of the sentiment coefficients overtime.

Sentiment scores were constructed from the first available frequency observation for each company to its last frequency observation. The number of frequency observations per company varies over the time frame and if no news is observed for a company at time t, the average term frequency, $f(v, i, t-1)$, and hence the sentiment score, $SCORE_{q,i,t}$, would be 0 at that moment.

3.1.3 Lexicon Selection

Given a set of coefficients, Λ_q, a sorting algorithm is proposed in order to select only relevant words for the specified quantile, q, and thus create an associated positive or negative lexicon. For example, one would intuitively expect a positive word such as *outperform* to be strongly associated with the upper quantiles (i.e. 90% quantile or 95% quantile) of returns while a negative word such as *slump* would be associated with lower quantiles (i.e 10% or 5%).

A sorting algorithm was constructed by fitting the coefficients using opposing quantiles to filter words. The objective was to filter out words that had opposite coefficient signs in opposite quantile regressions. Our hypothesis was that a word that has a *positive* or *negative* tone should maintain its sign regardless of the quantile. For example, a positive word such as *outperform* should be associated with an increase in the residual which is an indirect component of the company returns whether at the 5% quantile or 95% quantile. Therefore, $\lambda_{q,outperform}$ should be positive whether the quantile regression outlined in 3.1.2 is at the 95% quantile or the 5% quantile.

Sorting Algorithm

1. Perform the first regression using linear regression (equation 3.1) on the returns using the Fama-French factors as independent variables and obtain the residuals, $\eta_{i,t}$ for all firms.

2. For a given quantile, q, perform two quantile regressions on equation 3.2 in order to estimate the q-quantile and $(1-q)$-quantile of the residuals respectively.

3. Remove words which possess the opposite sign from both sets of quantile regression coefficients, $\Lambda_{q,unfiltered}$ and $\Lambda_{1-q,unfiltered}$ to obtain a new set of words, $V_{filtered}$.

4. Perform step 1 again and use quantile regression with Equation 3.2 on the obtained $V_{filtered}$ at desired quantile q to obtain a final set of coefficients, $\Lambda_{q,filtered}$, for vocabulary, $V_{filtered}$.

Thus, two sets of sentiment scores were constructed with vocabularies and coefficients, $\{V_{unfiltered}, \Lambda_{q,unfiltered}\}$ and $\{V_{filtered}, \Lambda_{q,filtered}\}$. Each company has a filtered and unfiltered sentiment score.

Score Cleaning

Average term frequencies can sometimes present extreme behaviour due to the surprise behaviour of the news/media cycle. To account for the *non-well behaved* nature of average term frequencies, the constructed scores were truncated within a specified range when used as regressors for the CAViaR models. First, the scores are truncated to be less than zero. Given that the residuals, $\eta_{i,t}$, are indirect components of company returns, the residuals should normally lower the mean return in equation 3.1 when accounting for the lower quantiles of returns (i.e. 10%, 5%, 1%). Second, for a given 5-year calibration window, the score observations are truncated to only contain observations above the 5% percentile within the window. This is used to re-

duce the effect of outliers, some of which are magnitudes larger than non-truncated observations and will heavily influence the CAViaR model.

For example, when fitting a CAViaR model for January 2006, sentiment observations between January 2000 and December 2005 above 0 are replaced by 0 and score observations below the empirical 5%-quantile of January 2000 to December 2005 data are replaced by the 5th quantile observation. Then, when applying the CAViaR model for the month of January 2006, the sentiment observations are truncated using 0 and the empirical 5%-quantile observed in the calibration window as cutoffs.

For each company, we thus have two sets, each with two series of sentiment scores. The first set is unbounded and the second set is bounded following truncation. They both contain two series, one obtained from the full set of words (unfiltered) and the other containing the filtered set of words (filtered). An example of the sentiment scores for Walmart at the 1% and 5% quantile is shown below (Figures 3.1, 3.2, 3.3, 3.4). The top-half per figure are the point observations and the bottom-half of the figure is the 20-day rolling average to distinguish the trend.

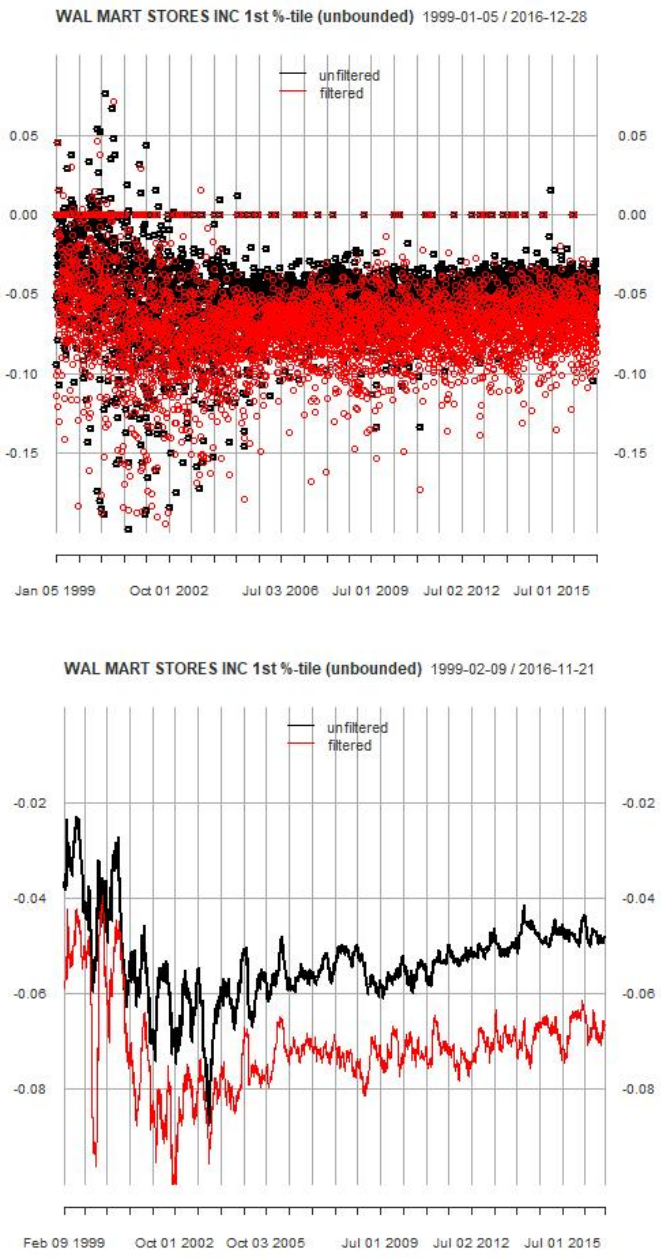

Figure 3.1: Walmart sentiment score (unfiltered versus filtered) at 1%-quantile (point observation and 20-day rolling average) **without** score truncation/cleaning (unbounded)

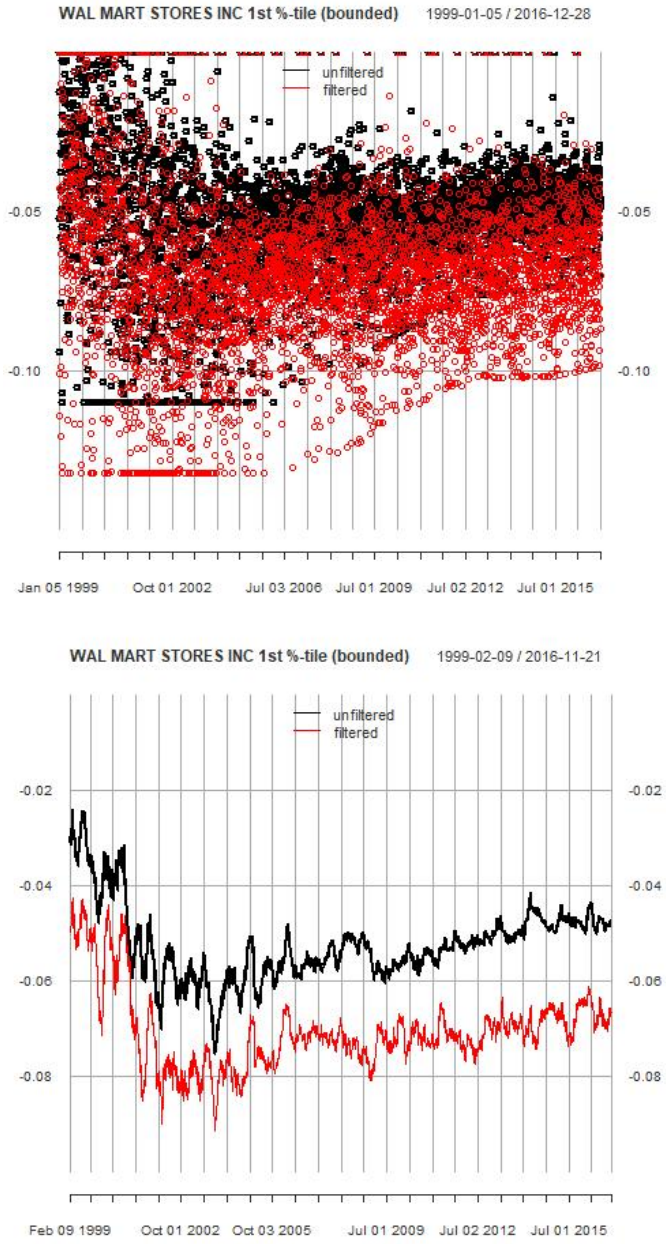

Figure 3.2: Walmart sentiment score (unfiltered versus filtered) at 1%-quantile (point observation and 20-day rolling average) **with** score truncation/cleaning (bounded)

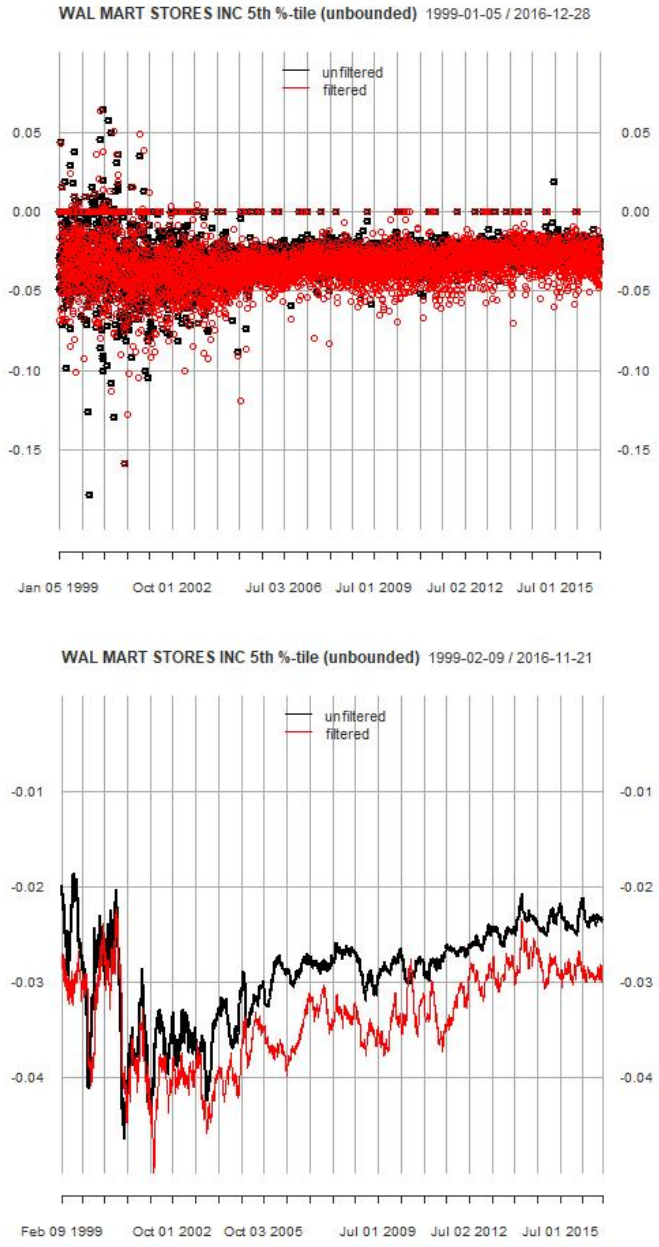

Figure 3.3: Walmart sentiment scores (unfiltered versus filtered) at 5%-quantile (point observation and 20-day rolling average) **without** score truncation/cleaning (unbounded)

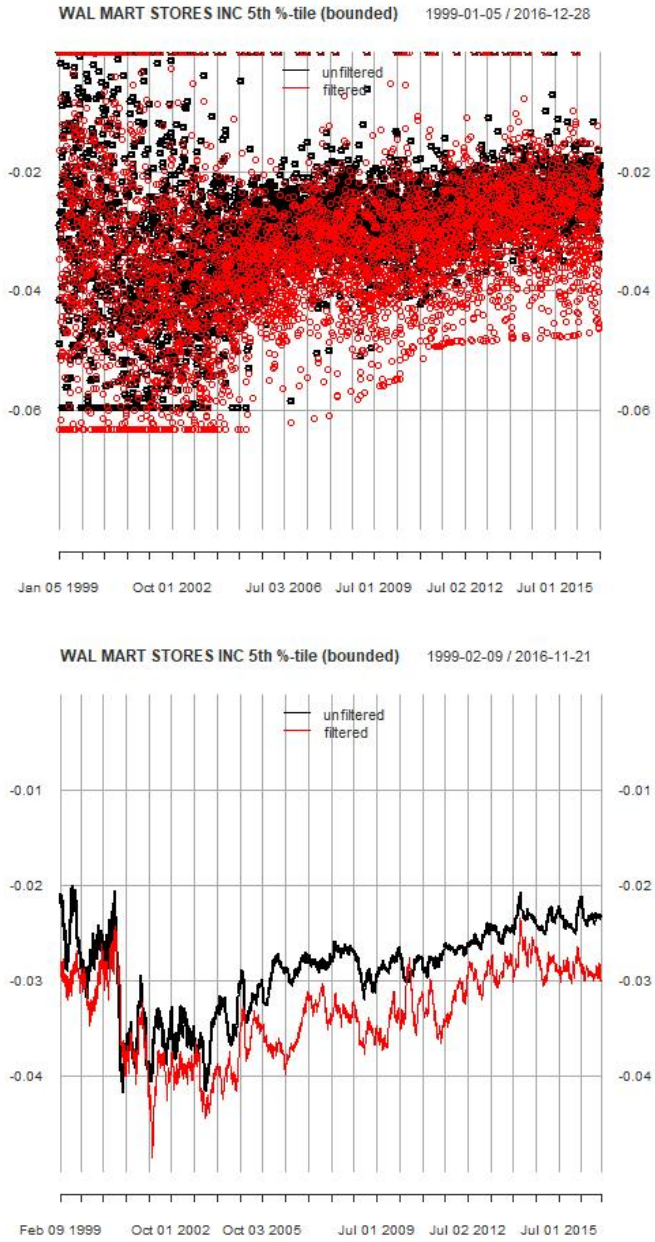

Figure 3.4: Walmart sentiment score (unfiltered versus filtered) at 5%-quantile (point observation and 20-day rolling average) **with** score truncation/cleaning (bounded)

3.2 CAViaR model calibration

To evaluate the additional predictive power for a set of coefficients, Λ_q, the following experiment is proposed:

1. For each company i, there is an associated series of returns, R_i and *one-day lagged* average term frequencies, $Freq_{i,t-1}$, for vocabulary V

2. The sentiment score for each company, $SCORE_{q,i,t}$ is calculated as outlined in equation 3.3 with the calibrated coefficients, Λ_q, for the entirety of companies in the dataset (section 2.3).

3. The sentiment score, $SCORE_{q,i,t}$, is used as an additional regressor in the following three CAViaR models alongside the returns:

   - **Augmented Symmetric Absolute Value:**

   $$f_t(\beta) = \beta_1 + \beta_2 f_{t-1}(\beta) + \beta_3 |r_{t-1}| + \beta_4 SCORE_{q,i,t} \tag{3.4}$$

   - **Augmented Asymmetric Slope:**

   $$f_t(\beta) = \beta_1 + \beta_2 f_{t-1}(\beta) + \beta_3 (r_{t-1})^+ + \beta_4 (r_{t-1})^- + \beta_5 SCORE_{q,i,t} \tag{3.5}$$

   - **Augmented Indirect GARCH(1,1):**

   $$f_t(\beta) = (\beta_1 + \beta_2 f_{t-1}^2(\beta) + \beta_3 r_{t-1}^2 + \beta_4 SCORE_{q,i,t}^2)^{\frac{1}{2}} \tag{3.6}$$

   These augmented models are compared to the benchmark models with no sentiment score regressor. The CAViaR models are calibrated on a monthly rolling basis using a 5-year window. For example, to calculate the VaR for the month of February 2005, the augmented and benchmark CAViaR models are calibrated using data from January 2000 to January 2005. Then, for the month of March 2005, the calibration window is February 2000 to February

2005. This is done for the whole time frame. A company with data from January 1999 to December 2016 would have 144 rolling periods. Given the computational intensity of this exercise, this was done for the top 100 companies out of 598 with the most average term frequency observations. This entails 144 periods $\times$ 100 companies = 14 400 CAViaR specifications per model per score. Parallel processing was used to calibrate the CAViaR specifications for individual companies. Table 3.1 outlines the model specifications per company per rolling period.

4. Obtaining 100 augmented VaR and 100 baseline series per specification, an aggregate performance comparison of the augmented models versus the original CAViaR models is done. The backtesting metrics include the DQ test, the AE ratio and the quantile loss ratio. The quantile loss ratio is:

$$\text{QL ratio} = \frac{\text{Augmented Model (CAViaRX or CAViaRfX)}}{\text{Baseline Model (CAViaR)}} \tag{3.7}$$

An example of the CAViaR models for Walmart at the 1% and 5% quantile is shown below (3.5, 3.6). For a given model specification (e.g. unbounded GARCH), the label CAViaR (in red) represents the baseline model, CAViaRX (in blue) represents the augmented CAViaR model with the unfiltered sentiment score and CAViaRfX (in green) represents the augmented CAViaR model with the filtered sentiment score.

Sentiment score per company ($\times$ 100)	
Unbounded values, unfiltered vocabulary	score type 1
Unbounded values, filtered vocabulary	score type 2
Bounded values, unfiltered vocabulary	score type 3
Bounded values, filtered vocabulary	score type 4
CAViaR model per company per rolling period ($\times$ 144)	
Baseline GARCH	CAViaR
GARCH with score type 1	Unbounded CAViaRX
GARCH with score type 2	Unbounded CAViaRfX
GARCH with score type 3	Bounded CAViaRX
GARCH with score type 4	Bounded CAViaRfX
Baseline SAV	CAViaR
SAV with score type 1	Unbounded CAViaRX
SAV with score type 2	Unbounded CAViaRfX
SAV with score type 3	Bounded CAViaRX
SAV with score type 4	Bounded CAViaRfX
Baseline AS	CAViaR
AS with score type 1	Unbounded CAViaRX
AS with score type 2	Unbounded CAViaRfX
AS with score type 3	Bounded CAViaRX
AS with score type 4	Bounded CAViaRfX

Table 3.1: Outline of model specifications per company per rolling period

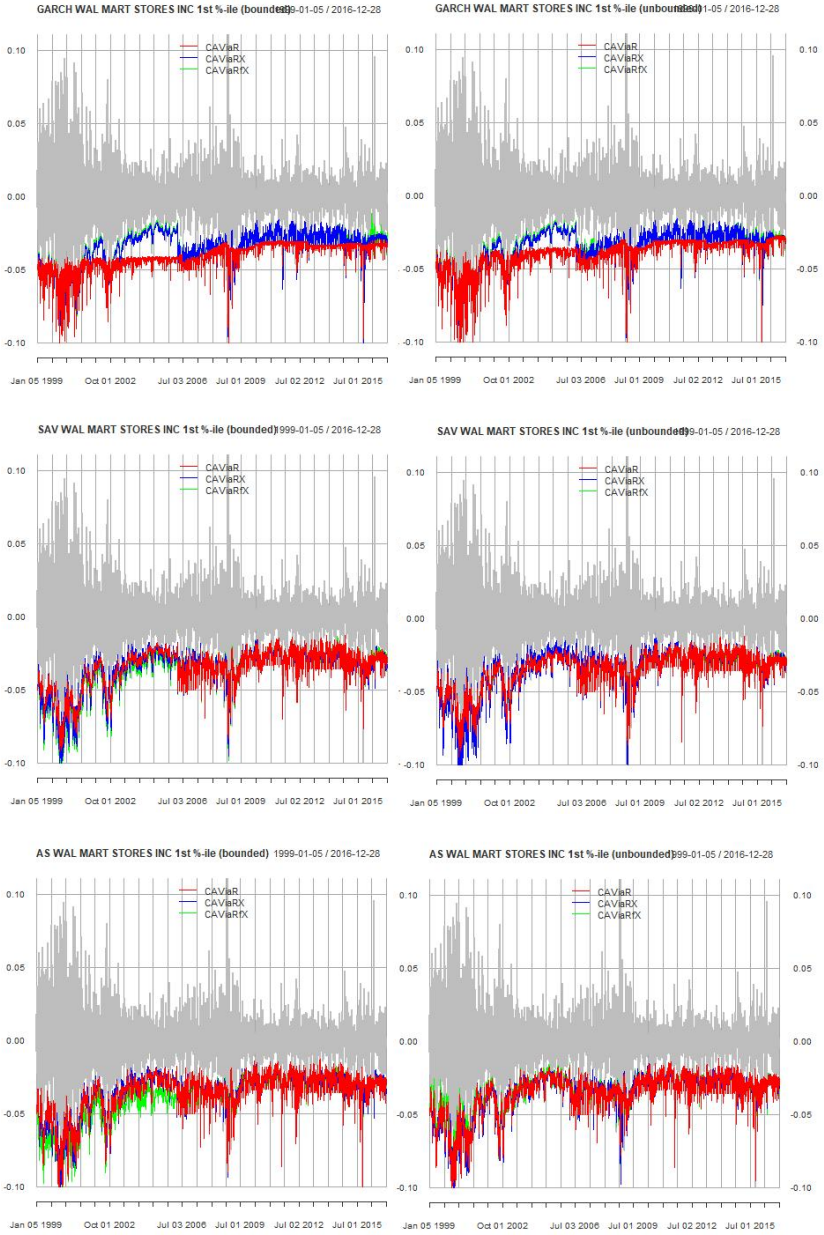

Figure 3.5: Walmart CAViaR models at the 1% quantile comparing bounded versus unbounded sentiment scores across the three CAViaR specifications (SAV, AS, GARCH)

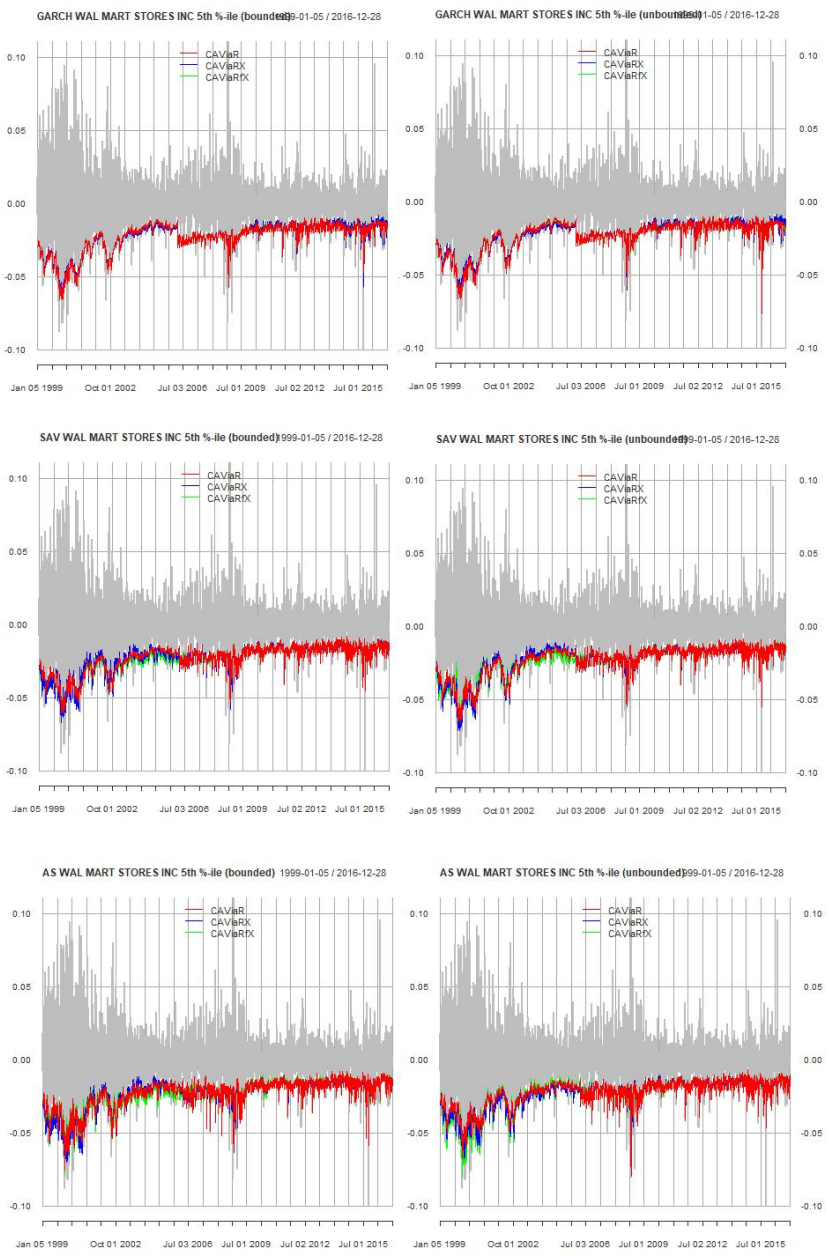

Figure 3.6: Walmart CAViaR models at the 5% quantile comparing bounded versus unbounded sentiment scores across the three CAViaR specifications (SAV, AS, GARCH)

Chapter 4

Empirical Results

4.1 Quantile Regression and Selection of Words

Upper and lower quantiles were arranged into pairs (i.e. 1% vs. 99% quantile regression and 5% vs. 95% quantile regression) and an initial regression was performed followed by a second filtering regression. The frequency coefficients were collected and the importance (A.1) of words in the vocabulary was observed. Importance is a standardized measure of coefficient variance as introduced by Manela & Moreira (2017). The sum of the importance across all words is equal to 100% by construction. Words with a high importance value account for large variability within the sentiment score. Table 4.1, 4.2, 4.3, and 4.4 show the results for the top 25 words with respect to importance for each subsequent regression and quantile level using the quantile regression technique. The cumulative importance of the top 25 words is relatively high (above 40%) meaning that they account for a high amount of variability within the score. Comparing the word selection of the low quantiles (i.e. 1% and 5% quantile) to the high quantiles (i.e. 95% and 99% quantile), we see that there is a large overlap of words with respect to importance. For example, Table 4.1, which represents the importance of words after the first quantile regression before filtering (1% vs. 99% quantile), the root words *contact, share, exclude* and *compani* are amongst the top 5 for both the regressions at the 1% quantile and at the 99% quantile. This could be explained due to a high variance in the average term

frequency of these words in the dataset. *Share* and *compani* would intuitively be commonly used words in financial literature.

The low quantile regressions however do manage to capture some negative words. In Table 4.1, words such as *unfit* (4th), *shortfal* (6th), *declin* (15th), *junk* (16th) and *slowdown* (25th) are amongst the top 25 importance words for the 1st quantile and are not present in the top 25 words for the corresponding high quantile regression at the 99th quantile. We notice however that in Table 4.3, first quantile regression before filtering (5th vs. 95th quantile), that these words become both present amongst the top 25 words.

Furthermore, the hypothesis that using opposing quantiles and opposite coefficient signs to filter words does not appear to hold much weight. When using the initial regression to filter out words with opposing signs and then perform a second regression, we find that almost all words in the top 25 for the initial regression are removed and that the top 25 words in the second regressions also appear to overlap as shown in Table 4.2, second filtering regression of 1st vs. 99th quantile, and Table 4.4, second filtering regression of 5th vs. 95th quantile.

Despite this, the second filtration does manage to carry over certain negative words. In Table 4.2, *unfit* (4th) is present in both the 1st regression and the 2nd subsequent filtration. In Table 4.2, other words including *disappoint* (15th) and *cautionari* (16th) are also amongst the top 25 words for the low quantile (1st) but not in the top 25 for the high quantile (99th). Comparing Table 4.3 to Table 4.4, more negative words appear to be singled out compared in the 1st quantile compared to the 5th.

The filtering process cuts a substantial amount of words in the vocabulary. From an initial 3585 words, the number of words remaining for the second regression in the 1%-vs-99% quantile regression is 1243 and 5%-vs-95% quantile regression is 1351 respectively. The cumulative importance of words is roughly the same (above 40%) amongst the top 25 words between the 1st and 2nd paired regressions.

A possible extension of this thesis would be to test other variable selection meth-

ods such as penalized quantile regression. As stated previously, researchers including Caporin & Poli (2017), Audrino *et al.* (2020) and Lehrer *et al.* (2021) use penalized regression methods (e.g. LASSO) to optimally select amongst a multitude of pre-constructed sentiment indices. The aim in this context would be to optimally select words amongst the initial amount of words in the vocabulary set.

		1% quantile				99% quantile	
		Lambda	Importance (%)			Lambda	Importance (%)
1.	*contact*	-3.5432	8.006	1.	*contact*	3.9748	9.566
2.	*share*	-1.6377	7.371	2.	*share*	1.7516	8.006
3.	*exclud*	-18.8684	5.420	3.	*compani*	0.9178	3.878
4.	*unfit*	-166.6197	4.832	4.	*loss*	2.6620	2.132
5.	*compani*	-0.77324	2.899	5.	*exclud*	11.9626	2.068
6.	*shortfal*	-20.6440	2.147	6.	*diminut*	140.4652	1.792
7.	*loss*	-1.9661	1.225	7.	*vice*	1.2853	1.392
8.	*lead*	-1.3870	1.185	8.	*lead*	1.4409	1.214
9.	*soft*	-22.4596	1.083	9.	*jitteri*	91.0328	0.946
10.	*vice*	-1.1025	1.079	10.	*call*	0.9596	0.926
11.	*call*	-1.0029	1.065	11.	*close*	1.3441	0.900
12.	*improprieti*	-43.0414	1.022	12.	*coincid*	54.8499	0.816
13.	*close*	-1.3695	0.984	13.	*mine*	2.5481	0.805
14.	*avail*	-3.6489	0.857	14.	*logic*	2.0829	0.765
15.	*declin*	-2.009	0.853	15.	*bailout*	19.736	0.735
16.	*junk*	-14.871	0.853	16.	*enabl*	1.655	0.735
17.	*actual*	-2.668	0.750	17.	*beset*	50.757	0.734
18.	*sane*	-68.990	0.697	18.	*sane*	72.307	0.727
19.	*logic*	-1.909	0.677	19.	*underreport*	63.275	0.659
20.	*consensu*	-4.084	0.650	20.	*monster*	2.462	0.649
21.	*unsold*	-26.422	0.637	21.	*avail*	3.127	0.598
22.	*well*	-1.066	0.620	22.	*gladden*	41.272	0.591
23.	*enabl*	-1.479	0.618	23.	*bankruptci*	2.946	0.585
24.	*crusad*	-25.867	0.535	24.	*differ*	2.778	0.581
25.	*slowdown*	-6.895	0.533	25.	*servic*	0.647	0.569
	Total:	-	46.524		**Total:**	-	42.367

Table 4.1: 1% vs 99% quantile - 1st quantile regression

		1% quantile				99% quantile	
		Lambda	Importance (%)			Lambda	Importance (%)
1.	*tax*	-4.609	5.212	1.	*gain*	9.991	8.070
2.	*strong*	-6.479	3.986	2.	*strong*	7.334	4.880
3.	*gain*	-6.650	3.741	3.	*tax*	3.998	3.748
4.	*unfit*	-205.649	3.100	4.	*opportun*	6.086	3.714
5.	*independ*	-5.403	3.003	5.	*independ*	5.240	2.700
6.	*weak*	-9.749	2.837	6.	*proprietari*	10.124	2.412
7.	*opportun*	-5.091	2.719	7.	*need*	5.808	2.310
8.	*proprietari*	-8.992	1.991	8.	*better*	5.657	1.948
9.	*need*	-4.988	1.783	9.	*reliabl*	5.877	1.825
10.	*save*	-4.633	1.528	10.	*save*	5.087	1.760
11.	*better*	-4.897	1.528	11.	*excess*	10.517	1.224
12.	*writeoff*	-32.084	1.466	12.	*confid*	7.184	1.195
13.	*reliabl*	-5.065	1.419	13.	*successfulli*	9.206	1.177
14.	*aggreg*	-7.504	1.360	14.	*aggreg*	7.096	1.162
15.	*disappoint*	-13.453	1.350	15.	*counsel*	5.111	1.030
16.	*cautionari*	-17.540	1.149	16.	*commit*	4.735	1.008
17.	*successfulli*	-8.803	1.126	17.	*strengthen*	7.676	0.960
18.	*miss*	-9.568	1.107	18.	*ensur*	5.610	0.914
19.	*confid*	-6.647	1.071	19.	*edg*	6.204	0.906
20.	*commit*	-4.701	1.040	20.	*safeti*	2.234	0.895
21.	*slow*	-9.092	1.026	21.	*weak*	5.533	0.873
22.	*premium*	-5.448	0.938	22.	*premium*	5.219	0.822
23.	*unanticip*	-23.646	0.907	23.	*limit*	7.408	0.762
24.	*limit*	-7.868	0.899	24.	*join*	6.365	0.753
25.	*ideal*	-9.093	0.862	25.	*learn*	5.491	0.696
	Total:	-	47.145		**Total:**	-	47.746

Table 4.2: 1% vs 99% quantile - 2nd quantile regression after filtering

		5% quantile				95% quantile	
		Lambda	Importance (%)			Lambda	Importance (%)
1.	*contact*	-1.821	12.099	1.	*contact*	2.075	13.622
2.	*share*	-0.623	6.102	2.	*share*	0.753	7.728
3.	*compani*	-0.396	4.357	3.	*compani*	0.437	4.596
4.	*loss*	-1.176	2.507	4.	*lead*	0.815	2.032
5.	*vice*	-0.651	2.152	5.	*exclud*	5.055	1.930
6.	*lead*	-0.745	1.956	6.	*loss*	1.050	1.735
7.	*exclud*	-4.153	1.503	7.	*vice*	0.624	1.712
8.	*logic*	-1.137	1.375	8.	*logic*	1.323	1.614
9.	*close*	-0.652	1.277	9.	*avail*	1.857	1.101
10.	*free*	-0.418	1.204	10.	*call*	0.458	1.101
11.	*well*	-0.608	1.153	11.	*enabl*	0.813	0.928
12.	*call*	-0.425	1.097	12.	*free*	0.391	0.910
13.	*mine*	-1.153	0.995	13.	*close*	0.578	0.870
14.	*declin*	-0.901	0.983	14.	*declin*	0.886	0.823
15.	*avail*	-1.629	0.978	15.	*differ*	1.439	0.814
16.	*slowdown*	-3.788	0.921	16.	*home*	0.409	0.797
17.	*enabl*	-0.699	0.791	17.	*well*	0.522	0.737
18.	*actual*	-1.143	0.788	18.	*mine*	1.009	0.660
19.	*home*	-0.3589	0.706	19.	*bankruptci*	1.3556	0.648
20.	*differ*	-1.219	0.674	20.	*divis*	0.684	0.617
21.	*common*	-0.581	0.661	21.	*downturn*	2.867	0.517
22.	*divis*	-0.630	0.604	22.	*eloqu*	19.671	0.505
23.	*downturn*	-2.822	0.578	23.	*involv*	1.732	0.499
24.	*servic*	-0.259	0.550	24.	*tire*	0.779	0.481
25.	*drop*	-0.997	0.516	25.	*major*	0.644	0.467
	Total:	-	46.527		**Total:**	-	47.447

Table 4.3: 5% vs 95% quantile - 1st quantile regression

		5% quantile				95% quantile	
		Lambda	Importance (%)			Lambda	Importance (%)
1.	*experi*	-3.031	5.571	1.	*experi*	3.340	6.359
2.	*effect*	-3.592	4.018	2.	*effect*	3.705	4.018
3.	*save*	-3.163	3.195	3.	*save*	3.395	3.461
4.	*reliabl*	-3.462	2.974	4.	*reliabl*	3.613	3.044
5.	*leadership*	-2.660	2.253	5.	*leadership*	2.788	2.326
6.	*premium*	-3.781	2.029	6.	*discontinu*	6.818	1.912
7.	*discontinu*	-6.715	1.973	7.	*premium*	3.768	1.894
8.	*weaker*	-11.334	1.811	8.	*robust*	6.659	1.656
9.	*light*	-2.553	1.668	9.	*light*	2.572	1.591
10.	*encourag*	-3.827	1.485	10.	*flexibl*	4.890	1.524
11.	*flexibl*	-4.543	1.399	11.	*compet*	5.431	1.448
12.	*legal*	-2.334	1.337	12.	*encourag*	3.856	1.417
13.	*liabil*	-4.108	1.266	13.	*approach*	3.796	1.390
14.	*cautionari*	-8.582	1.237	14.	*liabil*	4.326	1.320
15.	*robust*	-5.561	1.229	15.	*legal*	2.333	1.256
16.	*intellig*	-2.341	1.216	16.	*anomali*	12.317	1.239
17.	*anomali*	-11.708	1.191	17.	*particular*	4.802	1.199
18.	*approach*	-3.390	1.179	18.	*intellig*	2.358	1.160
19.	*accord*	-4.780	1.176	19.	*cautionari*	8.458	1.129
20.	*compet*	-4.733	1.170	20.	*hand*	3.520	1.118
21.	*satisfact*	-3.753	1.072	21.	*layoff*	2.622	1.106
22.	*connect*	-3.222	1.038	22.	*connect*	3.388	1.079
23.	*essenti*	-4.192	0.981	23.	*weaker*	8.960	1.064
24.	*ideal*	-4.511	0.953	24.	*essenti*	4.289	0.966
25.	*layoff*	-2.357	0.951	25.	*accord*	4.433	0.951
	Total:	-	44.372		**Total:**	-	45.629

Table 4.4: 5% vs 95% quantile - 2nd quantile regression after filtering

4.2 Distribution of Scores

Figure 4.1 and 4.2 show the distribution of the minimum, median and maximum values amongst the 100 sentiment score series with or without filtration and before or after bounding/truncation. Comparing the 1% quantile distributions to the 5% quantile distributions, the nature of the quantiles is apparent with the range of values for the 1% quantile to be lower than the range of values for the 5% quantile as is to be expected.

Second, comparing figure 4.1 against figure 4.2, the truncation has a substantial effect on the minimum and maximum values. Without the bounding, most maximum values are above zero and the minimums present more extremes.

Third, the filtering also shifts the distribution of scores to the left. This is seen by looking at the distribution of median scores in the unbounded scores (figure 4.1) and across the min, median and max score distributions in the bounded scores (figure 4.2). One possible reason for this might be that the filtering process singles out negative words though this has to be further investigated.

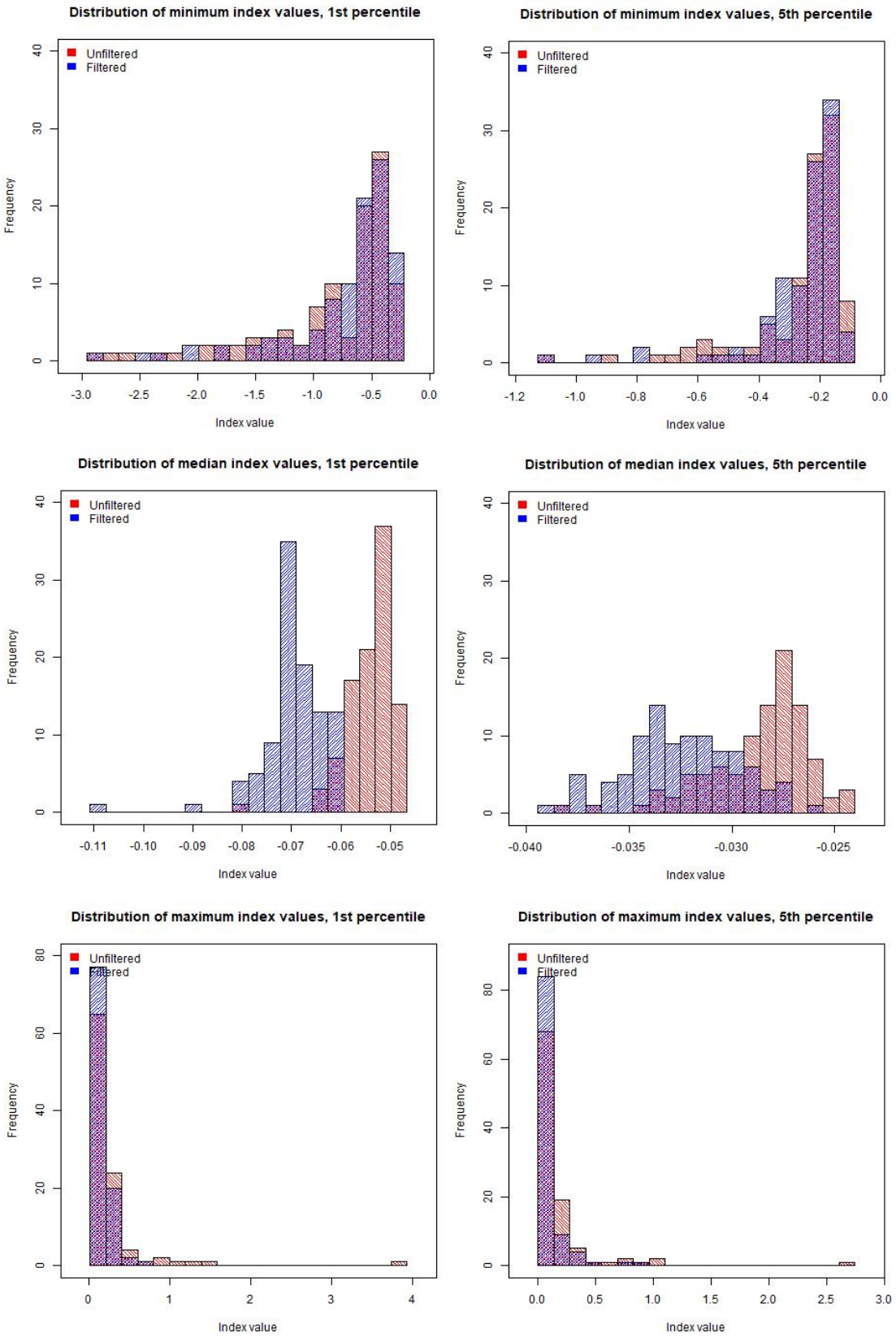

Figure 4.1: Distribution of sentiment score *unbounded* values for 100 companies

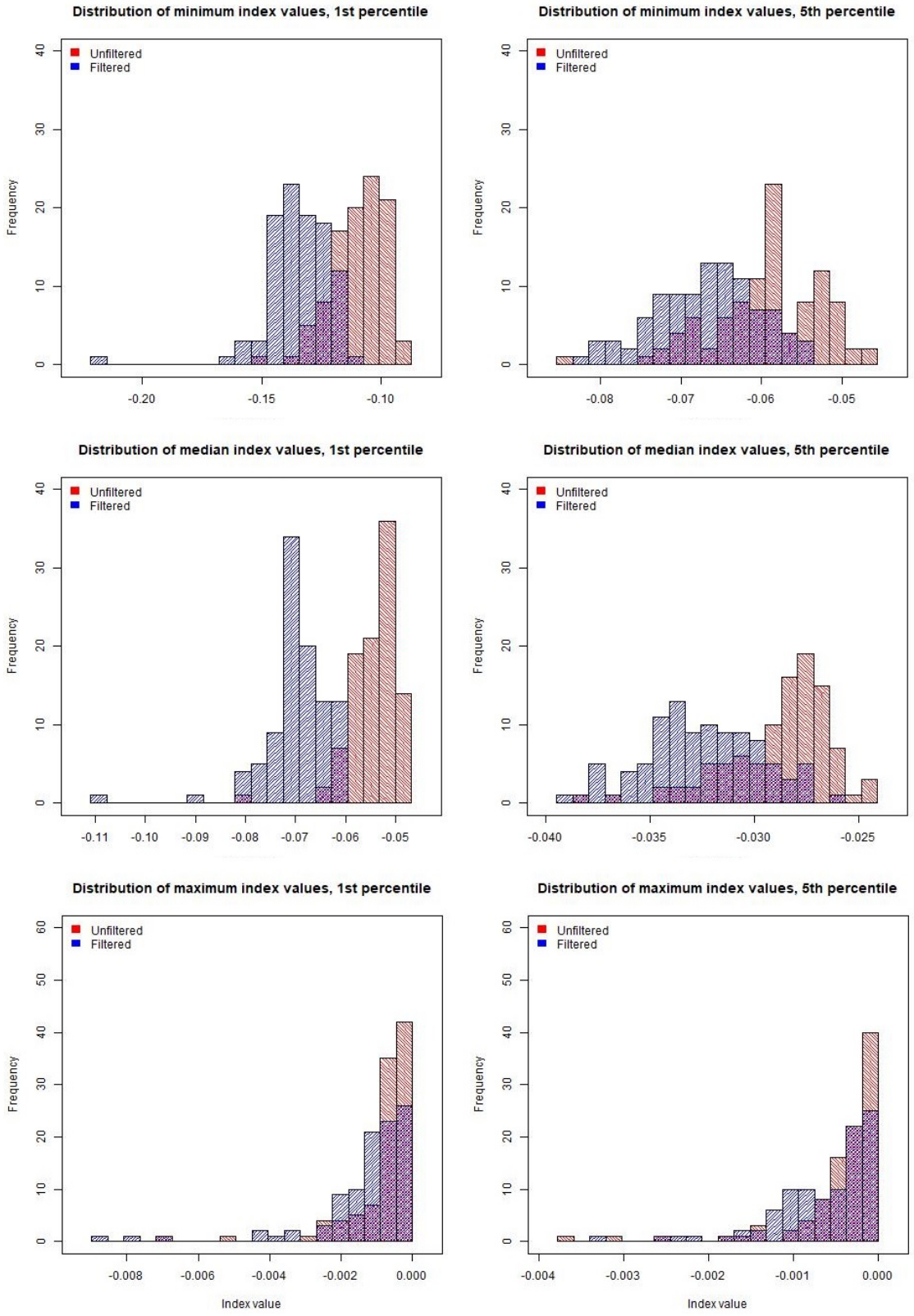

Figure 4.2: Distribution of sentiment score *bounded* values for 100 companies

4.3 Comparative Performance of CAViaR models

We evaluated the models using the DQ-test, the quantile loss measure and the actual-over-expected exceedance ratio (AE). The results for the three measures are outlined in Table 4.5.

4.3.1 DQ-Test

Panel A (Table 4.5) shows the number of DQ rejections out of the 100 companies at the 5% critical level across all model specifications. A smaller number suggests better model performance. Comparing the baseline model (CAViaR) to the augmented models (CAViaRX, CAViaRfX), at the 1% quantile, there is a noticeable improvement across all three CAViaR types (GARCH, SAV and AS) with less augmented models being rejected when the initial or filtered score is added. This trend is mixed at the 5% level.

For the baseline and augmented CAViaRs, the asymmetric slope models are the top performing group amongst the 3 major classes with the least number of rejections. At the 1% quantile score measure, the asymmetric slope (AS) models initially reject 39 out of 100 company calibrations but when the score is added as an additional regressor, the number of rejections drops to 21 (CAViaRX unbounded, CAViaRfX unbounded, CAViaRfX bounded) or 24 (CAViaRX bounded). Less rejections are also observed for the GARCH and Symmetric absolute value (SAV) models. However, this improvement is not apparent using the 5% quantile score measure. The number of regressions tends to slightly increase with the exception of CAViaRX bounded. One possible explanation to explain the 1% vs. 5% difference could be that sentiment tends to affect the more extreme nature in stock returns as previously studied by Garcia (2013), Ahmad *et al.* (2016), and Hanna *et al.* (2020).

There is some variation in the performance of the unbounded versus bounded groups, with the bounded group presenting less rejections for the GARCH and SAV models at the 1% quantile and GARCH at the 5% quantile. However, the difference is much less compared to baseline versus augmented models and this difference is

not found in the AS models (both 1% and 5% quantile) and the SAV model (at 5% quantile).

4.3.2 Quantile Loss Ratio

Panel B (Table 4.5) shows the number of models out of the 100 companies with a quantile loss ratio less than 1 compared to the baseline. The larger the number, the better the model performance. Overall, more than half (> 50) the models showed an improvement in quantile loss for most model variations (GARCH, SAV and AS at 1% quantile, GARCH and AS at 5% quantile). The SAV models at the 5% quantile improve for less than half of the 100 companies.

Comparing unbounded to bounded models, there is minimal difference in the number of improved models. The only noticeable changes are the AS models at the 1% quantile for CAViaRX and the GARCH at the 1% quantile for CAViaRfX where there is improvement from using the bounded score.

Looking between the 1% and 5% quantiles, the only noticeable improvement is the SAV model. For the SAV class of CAViaR models, the score calibrated at the 1% quantile improves more models compared to the score calibrated at the 5% quantile. For example, the loss ratio for the CAViaRX unbounded SAV models improves for 68 models compared to 48 models from the 5% to the 1% quantile. However, this improvement in performance between the 1% and 5% quantile is not observed with the GARCH or AS models.

Amongst the three model classes, more GARCH models improve with the addition of the exogenous variable compared to the SAV and AS models with the highest number of loss ratios below 1.

4.3.3 Joint DQ and Quantile Loss Ratio Criteria

Panel C (Table 4.5) shows the number of models out of 100 companies satisfying both the DQ test and a quantile loss ratio less than 1 compared to the baseline CAViaR. The larger the number, the better the model performance. The number

of models that satisfy both conditions falls to less than half in almost all cases with the exception of GARCH CAViaRX bounded at the 1st quantile at 55 models. Despite this, there is a clear differenciation in model performance looking between the 1% quantile and 5% quantile groups. There are more companies under the 1% quantile that satisfy both conditions compared to the 5% quantile calibration. This is apparent looking at the SAV and GARCH models between the 1% and 5% quantile. This improvement is minor for the AS models.

Comparing the unbounded and bounded scores, there is a minimal improvement using the bounded score with CAViaRX across GARCH, SAV and AS for both the 1% and 5% quantile. The result is mixed between the bounded and the unbounded scores when looking at CAViaRfX. A possible extension would be to investigate more companies or different VaR models (e.g. GARCH, HAR).

4.3.4 Actual over Exceedance Ratio

Panel D and panel E (Table 4.5) show the number of unbounded and bounded models respectively out of 100 companies with an AE ratio within a specified range. Two ranges of 0.98-1.02 and 0.95-1.05 are used. The larger the number, the better the model performance. Obviously, the larger range will contain the smaller range and have more models. Instead, the objective is to compare if there are noticeable increases for augmented models.

In both panels, comparing the baseline to the augmented models, the results are mixed. For example, the unbounded AS models (panel D) outperform the baseline CAViaR but this is not apparent in the bounded AS models (panel E).

In both panels, comparing the 1% and 5% quantile, we see that more models fall in the AE range at the 5% quantile than the 1% model for most model specifications. Exceptions to this trend are AS CAViaRfX at the 0.98 - 1.02 range in panel D (19 to 13) and SAV CAViaRfX at the 0.98 - 1.02 range in panel E (14 to 9). This observation is opposite to our analysis of the DQ and quantile loss ratio.

It is important to note that the AE ratio only tests the unconditional coverage

hypothesis, 1.22 while the joint DQ and Loss Ratio results from panel C would test both the unconditional coverage and independence hypothesis.

Table 4.5: Backtesting aggregate results for all CAViaR models

Panel A:

Number of DQ rejections out of 100 companies at 5% critical level

	1% quantile			5% quantile		
	GARCH	SAV	AS	GARCH	SAV	AS
CAViaR	52	44	39	79	57	30
CAViaRX (*unbounded*)	43	38	21	79	60	33
CAViaRX (*bounded*)	35	30	24	76	56	26
CAViaRfX (*unbounded*)	41	35	21	80	57	34
CAViaRfX (*bounded*)	31	32	21	78	66	37

Panel B: Number of Loss Ratios < 1 out of 100

	1% quantile			5% quantile		
	GARCH	SAV	AS	GARCH	SAV	AS
CAViaRX (*unbounded*)	74	68	57	74	48	63
CAViaRX (*bounded*)	76	69	69	78	43	66
CAViaRfX (*unbounded*)	68	60	58	67	38	52
CAViaRfX (*bounded*)	75	62	56	69	38	54

Panel C: Number of models satisfying both conditions out of 100

	1% quantile			5% quantile		
	GARCH	SAV	AS	GARCH	SAV	AS
CAViaRX (*unbounded*)	48	46	49	13	19	47
CAViaRX (*bounded*)	55	49	54	19	20	52
CAViaRfX (*unbounded*)	46	41	46	12	19	36
CAViaRfX (*bounded*)	53	44	45	14	15	34

Panel D: Number of models within specified AE range (unbounded)

	1% quantile			5% quantile		
	GARCH	SAV	AS	GARCH	SAV	AS
0.98 - 1.02						
CAViaR	9	8	8	15	12	11
CAViaRX	8	9	11	17	15	15
CAViaRfX	9	7	19	23	16	13
0.95 - 1.05						
CAViaR	23	22	22	45	34	29
CAViaRX	28	23	25	44	35	42
CAViaRfX	33	23	26	47	35	33

Panel E: Number of models within specified AE range (bounded)

	1% quantile			5% quantile		
	GARCH	SAV	AS	GARCH	SAV	AS
0.98 - 1.02						
CAViaR	11	9	8	22	10	14
CAViaRX	11	6	11	21	15	17
CAViaRfX	8	14	7	22	9	12
0.95 - 1.05						
CAViaR	26	19	23	40	23	30
CAViaRX	20	21	29	47	36	38
CAViaRfX	21	28	20	48	28	34

Chapter 5

Conclusion

The use of non-conventional sources of data including textual sentiment continues to grow in importance especially in applications such as economics and finance and risk-management. Thus far, quantile regression is a novel technique in its use for natural language processing and specifically, textual sentiment analysis.

By proposing a score derived from quantile regression on textual data concerning non-financial firms, we demonstrated that there is a marginal improvement in VaR backtesting performance for CAViaR models at the 1% quantile level. This conclusion is in sync with the idea that sentiment is associated with market reactions and extreme idiosyncratic events. Possible extensions of this research could include reformulating the quantile regression such that the idiosyncratic element of returns is obtained otherwise (e.g. definition of an abnormal return versus the systemic market return) or implementing other variable selection methods to select the lexicon of choice words (e.g. penalized quantile regression).

Appendix A

Importance of Words

Importance is a standardized measure of coefficient variance as introduced by Manela & Moreira (2017). It is constructed using the product of squared regression coefficients and term-frequency variance. The sum of the importance across all words is equal to 100% by construction.

For a set of term-frequency regression coefficients for V words, $\{\lambda_v\}_{v=1}^{V}$, the importance for word v is defined as:

$$imp_v = \frac{\hat{\lambda_v^2}\hat{\sigma_v^2}}{\sum_{j=1}^{J}\hat{\lambda_v^2}\hat{\sigma_v^2}} \tag{A.1}$$

where:

- $\hat{\lambda_v}$ is the coefficient estimate of word v
- $\hat{\sigma_v^2}$ is the variance of the term-frequencies for word v

References

1. Ahmad, K., Han, J., Hutson, E., Kearney, C. & Liu, S. Media-expressed Negative Tone and Firm-level Stock Returns. *Journal of Corporate Finance* **37,** 152–172 (2016).

2. Algaba, A., Ardia, D., Bluteau, K., Borms, S. & Boudt, K. Econometrics Meets Sentiment: An Overview of Methodology and Applications. *Journal of Economic Surveys* **34,** 512–547 (2020).

3. Allen, D. E., McAleer, M. & da Veiga, B. Modelling and Forecasting Dynamic VaR Thresholds for Risk Management and Regulation. *Available at SSRN 926270* (2005).

4. Antweiler, W. & Frank, M. Z. Is All that Talk just Noise? The Information Content of Internet Stock Message Boards. *Journal of Finance* **59,** 1259–1294 (2004).

5. Ardia, D., Bluteau, K. & Boudt, K. Questioning the News about Economic Growth: Sparse Forecasting using Thousands of News-based Sentiment Values. *International Journal of Forecasting* **35,** 1370–1386 (2019).

6. Ardia, D., Boudt, K. & Catania, L. Generalized Autoregressive Score Models in R: The GAS Package. *Journal of Statistical Software* **88,** 1–28 (2019).

7. Ardia, D., Boudt, K. & Catania, L. Value-at-Risk Prediction in R with the GAS Package. *R Journal* **10,** 410–421 (2018).

8. Audrino, F., Sigrist, F. & Ballinari, D. The Impact of Sentiment and Attention Measures on Stock Market Volatility. *International Journal of Forecasting* **36,** 334–357 (2020).

9. Baker, M. & Wurgler, J. Investor Sentiment in the Stock market. *Journal of Economic Perspectives* **21,** 129–152 (2007).

10. Banerjee, A. K., Dionisio, A., Pradhan, H. & Mahapatra, B. Hunting the Quicksilver: Using Textual News and Causality Analysis to Predict Market Volatility. *International Review of Financial Analysis* **77,** 101848 (2021).

11. Boudoukh, J., Feldman, R., Kogan, S. & Richardson, M. *Which News Moves Stock Prices? A Textual Analysis* tech. rep. (National Bureau of Economic Research, 2013).

12. Bradley, M. M. & Lang, P. J. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings* tech. rep. (Technical report C-1, the Center for Research in Psychophysiology ..., 1999).

13. Caporin, M. & Poli, F. Building News Measures from Textual Data and an Application to Volatility Forecasting. *Econometrics* **5,** 35 (2017).

14. Chan, W. S. Stock Price Reaction to News and No-News: Drift and Reversal after Headlines. *Journal of Financial Economics* **70,** 223–260 (2003).

15. Christoffersen, P. F. Evaluating Interval Forecasts. *International Economic Review* **39,** 841–862 (1998).

16. Da, Z., Engelberg, J. & Gao, P. In Search of Attention. *Journal of Finance* **66,** 1461–1499 (2011).

17. Damodaran, A. Value at risk (VAR). *Stern School of Business at New York University, abs/1308.2066* (2007).

18. Das, S. R. & Chen, M. Y. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science* **53,** 1375–1388 (2007).

19. Dumitrescu, E.-I., Hurlin, C. & Pham, V. Backtesting Value-at-Risk: from Dynamic Quantile to Dynamic Binary Tests. *Finance* **33,** 79–112 (2012).

20. Engle, R. F. & Manganelli, S. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics* **22,** 367–381 (2004).

21. Fama, E. F. & French, K. R. A Five-Factor Asset Pricing Model. *Journal of Financial Economics* **116,** 1–22 (2015).

22. Fama, E. F. & French, K. R. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* **33,** 3–56 (1993).

23. French, K. R. *Data library* Tuck School of Business at Dartmouth faculty web profile for Kenneth R. French, available at http://mba. tuck. dartmouth. edu/pages/faculty/ken. french/data_library. html. 2013.

24. Garcia, D. Sentiment During Recessions. *Journal of Finance* **68,** 1267–1300 (2013).

25. Gentzkow, M., Kelly, B. & Taddy, M. Text as Data. *Journal of Economic Literature* **57,** 535–74 (2019).

26. González-Rivera, G., Lee, T.-H. & Mishra, S. Forecasting Volatility: A Reality Check based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood. *International Journal of Forecasting* **20,** 629–645 (2004).

27. Hanna, A. J., Turner, J. D. & Walker, C. B. News Media and Investor Sentiment during Bull and Bear Markets. *European Journal of Finance* **26,** 1377–1395 (2020).

28. Henry, E. Are Investors Influenced by how Earnings Press Releases are Written? *Journal of Business Communication* **45,** 363–407 (2008).

29. Heston, S. L. & Sinha, N. R. News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal* **73,** 67–83 (2017).

30. Jegadeesh, N. & Wu, D. Word Power: A New Approach for Content Analysis. *Journal of Financial Economics* **110,** 712–729 (2013).

31. Jeon, J. & Taylor, J. W. Using CAViaR Models with Implied Volatility for Value-at-Risk Estimation. *Journal of Forecasting* **32,** 62–74 (2013).

32. Jones, K. S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* **28,** 11–21 (1972).

33. Keynes, J. M. *The General Theory of Interest, Employment and Money* (London: MacMillan, 1936).

34. Koenker, R. & Bassett Jr, G. Regression Quantiles. *Econometrica: Journal of the Econometric Society* **46,** 33–50 (1978).

35. Koenker, R., Portnoy, S., *et al.* Package 'quantreg'. *Cran R-project. org* (2018).

36. Kothari, S. P., Li, X. & Short, J. E. The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study using Content Analysis. *Accounting Review* **84,** 1639–1670 (2009).

37. Kupiec, P. Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives* **3,** 73–84 (1995).

38. Lehrer, S., Xie, T. & Zhang, X. Social Media Sentiment, Model Uncertainty, and Volatility Forecasting. *Economic Modelling* **102,** 105556 (2021).

39. Liu, B. *et al.* Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing* **2,** 627–666 (2010).

40. Loughran, T. & McDonald, B. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research* **54,** 1187–1230 (2016).

41. Loughran, T. & McDonald, B. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance* **66,** 35–65 (2011).

42. Luhn, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* **2,** 159–165 (1958).

43. Manela, A. & Moreira, A. News Implied Volatility and Disaster Concerns. *Journal of Financial Economics* **123,** 137–162 (2017).

44. Manning, C. & Schutze, H. *Foundations of Statistical Natural Language Processing* (MIT press, 1999).

45. Picault, M. & Renault, T. Words are not all Created Equal: A New Measure of ECB Communication. *Journal of International Money and Finance* **79,** 136–156 (2017).

46. Portnoy, S. & Koenker, R. The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators. *Statistical Science* **12,** 279–300 (1997).

47. Shiller, R. J. *Irrational Exuberance* (Princeton university press, 2000).

48. Shiller, R. J. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events* (Princeton University Press, 2020).

49. Stone, P. J. & Hunt, E. B. *A Computer Approach to Content Analysis: Studies using the General Inquirer System* in *Proceedings of the May 21-23, 1963, Spring joint computer conference* (1963), 241–256.

50. Tetlock, P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance* **62,** 1139–1168 (2007).

51. Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. More than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance* **63,** 1437–1467 (2008).

52. Wisniewski, T. P. & Lambe, B. The Role of Media in the Credit Crunch: The Case of the Banking Sector. *Journal of Economic Behavior & Organization* **85,** 163–175 (2013).