Properties of Financial Texts

Vitaliy Ryabinin[1]
Imperial College Business School
v.ryabinin19@imperial.ac.uk

Abstract

Statistical properties of unstructured data are largely unknown. I find that counts of words (positive, negative, text length), their combinations, and measures constructed from them are often non-stationary. For most of these time series, the ADF test rejects the null hypothesis of unit root presence. On the other hand, the KPSS test rejects trend stationarity. Visual evidence aligns with the KPSS outcome. This pattern is more pronounced for daily data. A direct comparison between conventional frequency-based measure of news sentiment and a stationary counterpart demonstrates the economic impact. Predicting market returns with a non-stationary word frequency measure results in contradictory empirical findings. Forecast errors and prediction beta are higher in recessions than expansions at the same time. After accounting for the stationarity, the magnitude of beta decreases by over 50%, implying that the sentiment's influence on the equity market returns has been severely overstated.

Keywords: News Sentiment, Stationarity, Unstructured Data, Predictability, Model Risk.

JEL Classification: C5; E32.

[1]Imperial College Business School, South Kensington Campus. 14 Prince's Gardens, Floor 4, Room 14.403, London SW7 1NA, United Kingdom.

1 Introduction

Are economic conclusions negatively affected by the unknown statistical properties of unstructured data? Numeric representations of textual features such as word counts (positive, negative, text length) are often non-stationary. Combinations of them (ratio of relevant words to text length, difference between positive and negative terms, etc.) are non-stationary as well. All these are basic inputs into variable construction algorithms. Subsequently, the resulting measures inherit properties of the underlying components. Accounting for the non-stationarity of news-based sentiment demonstrates the ramifications. Contrary to prior studies, both market return and trading volume predictions have higher forecast errors in recessions.

Unstructured data, such as text and images, gain an ever-increasing prominence in economic variable construction, mainly due to high frequency and low cost. The upside is availability, using text broadens the range of feasible empirical applications. The downside is clear as well; there is a higher likelihood of flawed conclusions. The problem is further exacerbated by the opacity of common measure construction procedures. Often enough, the underlying components (such as individual word counts) are unavailable. Without assessing the inputs, it is impossible to tell whether the output is accurate or is due to a spurious relationship.[2] As a result, establishing common statistical properties of the unstructured data and incorporating robust inference at least partially mitigates the issue.

Employing robust inference provides an additional, perhaps most important, safety layer. Regression betas, which are by far the most common measures, are strongly impacted by non-stationarity, especially in small samples. On the other hand, statistics unaffected by stationarity, such as mean absolute (squared) error, yield reliable conclusions. For example, Garcia (2013) uses multiple different term frequencies (counts of positive and negative words divided by text length) to proxy for the news-based sentiment. The measures are non-stationary. The outcome of his analysis, "the

[2]In machine learning, this concept is often referred to as "GIGO" - garbage in, garbage out.

predictability of stock returns using news' content is concentrated in recessions", is based on relative magnitudes of regression coefficients. However, as demonstrated later in this paper, the forecast errors (based on the exact same measures) are higher in recessions. As a result, unknown and undesirable statistical properties (such as non-stationarity) propagate through the estimation methods and drastically affect the resulting economic conclusions.

Machine learning is commonly used in conjunction with the alternative data. The inputs are rarely preprocessed and much more often are just "tossed" into a black box. However, both basic and advanced learning algorithms are not devoid of assuming stationarity, either explicitly or implicitly. For example, all models augmented with the regularization or variable selection (such as LASSO, Ridge, Elastic net, etc.), directly carry it over from the ordinary least squares. Alternatively, other machine learning methods assume that the data generation process does not change over time. Latent Dirichlet Allocation (Blei et al. (2003)) views documents as coming from a random mixture over latent features (topics) with a fixed word distribution. Neural networks follow suit. The network itself and its set of parameters are usually fixed and do not change over time. While the machine learning methods might still provide a good in-sample fit under non-stationarity, it is unlikely to carry over out-of-sample. However, the importance of stationarity for the actual real-world performance depends on the method and its intended application. Gentzkow et al. (2019) provide a comprehensive overview of the text processing algorithms, underlying assumptions, and usage cases.

From this perspective, cross-validation is especially difficult in a time-series setting under non-stationarity. It is very hard to split a series into representative training and testing samples as their properties would not be consistent. Finally, the model fit is often evaluated using the R^2 (or its derivatives), which may be affected by a spurious relationship. Robust alternatives, such as mean absolute error, aren't always adaptable to the application at hand. Covariate shift (the distribution of

inputs post-validation differs from that of testing and training samples) is perhaps the best specific example. It is hard to tell if a trickle of post-validation data arriving in real time falls within a stress-testing window. This problem is pronounced in forecasting applications. New York Fed's nowcasting was indefinitely suspended due to unexpectedly extreme COVID data.[3] It is telling that the issue befell the New York Fed; the Federal Reserve system is responsible for supplying[4] a comprehensive set of recommendations and best practices on the model risk management. In this light, less regulated or structured industries should be affected even more.

Consequently, it is important to pre-test both the inputs (if available) and a constructed measure. I check the stationarity of multiple prominent measures, all based on the unstructured data, in order to demonstrate that there are common statistical properties, independent from the intended applications. First, the testing procedure matters greatly. While multiple stationarity tests are available, the results do not necessarily agree. Given unstructured data or a measure constructed from it, with a very few exceptions, the augmented Dickey-Fuller (ADF, Dickey & Fuller (1979)) test rejects the null hypothesis of unit root presence. On the other hand, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS, Kwiatkowski et al. (1992)) test rejects trend stationarity. Visual evidence aligns with the KPSS outcome. This situation is quite rare compared to the conventional, structured data. For example, for a set of fourteen economic variables in Kwiatkowski et al. (1992), only the industrial production series has "evidence against both hypotheses". It is followed by a suggestion to consider "other alternatives, such as explosive roots, fractional integration, or stationarity around a nonlinear trend".

Based on the Johansen (1991) test, relevant term counts and text lengths are cointegrated. This property is likely a characteristic of both natural language and a

[3] Federal Reserve Bank of New York, "The uncertainty around the pandemic and the consequent volatility in the data have posed a number of challenges to the Nowcast model. Therefore, we have decided to suspend the publication of the Nowcast while we continue to work on methodological improvements to better address these challenges".

[4] SR Letter 11-7; "Supervisory Guidance on Model Risk Management".

writing process. As the number of words in a text (length) goes up, so does the count of positive and negative words. It should hold for a variety of applications, unless the text is broken up into sections, each dedicated to a narrow topic written in specialist jargon. Investigating further which specific documents may be characterized as such is outside the scope of this paper, but Item 1A "Risk Factors" section of the 10-K filings, on the surface, satisfies the criteria.

If word counts and text lengths are available, then the Engle & Granger (1987) two-step method is directly applicable and yields a stationary measure. Unlike a simple ratio of word count to text length, the procedure accounts for non-stationarity. However, it is not a "free lunch"; performing econometrically correct inference becomes more complex. The two-step method results in an imputed regressor measured with a sampling error (Murphy & Topel (1985)). Such variables render inapplicable commonly used covariance adjustments such as those described in White (1980) and Newey & West (1987). The inference issue is addressed here in two separate ways. First, a highly robust Ibragimov & Muller (2010) test statistic accounts for the data imperfections by relying on the heavy tails of Student's t-distribution with a low number of degrees of freedom. Second, substituting quantile regression for the OLS sidesteps the problem altogether. Quantile regression does not have a closed form solution so the inference does not rely on covariance adjustments. Finally, the quantile regression is also generally less sensitive to outliers.

The relationship between news-based sentiment and financial markets is explored further by employing a stationary measure constructed using the Engle & Granger (1987) two-step method. To make the results immediately comparable to prior studies, the stationary news-based sentiment measure relies on word counts from Garcia (2013). It is used to forecast daily market returns and trading volume. At best, the measure predicts about 4 basis points of the daily Dow returns. The estimate is statistically significant but economically inconsequential. Harvesting the prediction premium prior to the Internet would have been nearly impossible. Calling a broker

(or faxing an order) involved high fees and a significant front running risk. On the other hand, the link between aggregated trading volume and news-based sentiment is both economically and statistically significant. The relationship is inverse; as negative sentiment rises by one standard deviation, trading volume falls by 1.5 percentage points. Additionally, trading volume predictions have higher R^2 values than the return forecasts suggesting that the market and news are connected, just not through prices. These findings, at least partially, validate using trading volume as one of the proxies for sentiment, as suggested in M. Baker & Wurgler (2007). Overall, a stronger link between the sentiment and trading volume is consistent with the disagreement models (Hong & Stein (2007)) where difference in opinions causes price signals to simultaneously cancel each other and aggregate to a higher number of transactions.

2 Properties of prominent text-based measures

2.1 Text-based measures and their applications

Text is used for the creation of both long (spanning decades) and high frequency (daily) variables. These measures often have no substitutes making it harder to verify the economic conclusions. To that extent, I test multiple time series based on the unstructured data to determine if there is a connection between construction algorithm, topic, and resulting statistical properties. The focus is on stationarity as it directly affects inference and economic interpretation of the regression coefficients.

For the majority of variable created from the unstructured data, the ADF test rejects the null hypothesis of unit root presence while the KPSS rejects trend-stationarity. In general, this is a fairly rare scenario. For example, it was encountered only for the industrial production time series (out of 14 total) during the empirical validation of KPSS test (Kwiatkowski et al. (1992)). Kwiatkowski et al. (1992) did not resolve the ensuing ambiguity writing "there is evidence against both hypotheses, and thus it is not clear what to conclude".

As a result, to establish stationarity,[5] the KPSS and ADF tests need to be considered jointly. Any outcome other than simultaneously rejecting the null (ADF) and failing to reject the null (KPSS) indicates a potential issue. If necessary, visual evidence acts as a tiebreaker. Based on these criteria, most of the measures in (Table 1) are non-stationary. The stationarity does not appear to be dependent on the topic or construction method but may be related to frequency (Table 1). Daily measures created from the unstructured data are almost guaranteed to be non-stationary. The economic applications of all tested measures are briefly discussed below.

The main result of Garcia (2013) is that "the predictability of stock returns using news' content is concentrated in recessions". The conclusion is obtained by comparing the magnitudes of regression coefficients. Obaid & Pukthuanthong (2021) adopts Garcia (2013) regression specification and concludes, "Photo Pessimism predicts market return reversals and trading volume". Shapiro et al. (2020) find that "text-based measures of sentiment extracted from news articles perform well in terms of capturing economically meaningful soft information". S. R. Baker et al. (2016) rely on the regression coefficients to validate their text-based measure. Specifically, "for every 1% increase in our policy uncertainty index a firm with, say, a 50% government revenue share would see its stock volatility rise by 0.11%". Additionally, S. R. Baker et al. (2016) economic policy uncertainty measure is a commonly used control. For example, Manela & Moreira (2017) construct a text-based measure of implied volatility that "predicts disasters" and also use S. R. Baker et al. (2016) time series as one of the controls. Bybee et al. (2020), Caldara et al. (2020), Caldara & Iacoviello (2022) create numerous time series from the unstructured data and use them to draw industry and firm level conclusions. Economic importance, in most of these cases, depends on the magnitude and significance of the coefficient of interest.

[5] Conditional on satisfying the underlying assumptions, mainly linearity.

2.2 Statistical properties of word counts

The focus of subsequent analysis is on news-based sentiment. The findings, however, are more general. News articles on economic activity are a common data source. Intrinsic properties of texts such as lengths are independent from the topic. Sentiment measurement has a long history of relying on both text and conventional, structured data. As a result, it is possible to assess the validity of testing procedure, properties of the measure components, and the connection to financial markets. The data[6] is from Garcia (2013). It includes three daily series spanning 1905-2005: counts of positive words, negative words, and text lengths. They are sourced from two New York Times columns, "Financial Markets" and "Topics in Wall Street", then classified using the Loughran & McDonald (2011) dictionaries.

The ADF and KPSS tests are again discordant (Table 2) for both word counts and frequencies (Pos/Len, Neg/Len, and $(Neg - Pos)/Len$). The ADF test indicates trend-stationarity, while the KPSS rejects it.[7] Daily Dow Jones returns are included to demonstrate that a known stationary variable passes both hurdles. Visual evidence (Figure 1) supports the KPSS test[8] outcome. The yearly average is changing over time for all word counts and frequencies. Trend-stationarity requires the immutability of the unconditional joint probability distribution with respect to a deterministic trend. The yearly mean is not stable, there is no deterministic trend, and fluctuations do not have a discernible pattern.

There is a resemblance between the word counts and text lengths (Figure 1) hinting at cointegration. The individual measure components are also highly correlated with the text lengths (Table 4). For example, the count of positive word and text lengths have a correlation coefficient of 0.84. Following Gonzalo & Lee (1998) and Haug (1996), the cointegration is formally tested using the Johansen (1991) proce-

[6] Obtained from Garcia's website.

[7] The results (both ADF and KPSS) are robust to the number of lags.

[8] Stationarity is rarely explicitly tested, and even then the KPSS test is often overlooked. For example, Kalamara et al. (2022) exclusively use the ADF test. Moreover, common textbooks do not mention the additional tests. For example, Stock & Watson (2019) only includes the ADF test.

dure. Test statistics for all possible pairwise combinations of word counts greatly exceed critical values (Table 3) indicating an underlying cointegrating relationship. Concerns regarding the Johansen test specification, "the Johansen estimator finds too much cointegration when the lag order is misspecified, which may lead researchers to examine long-run relations which are actually spurious" (Ho & Sorensen (1996)), are likely irrelevant for the application at hand. The data has 27,447 non-missing daily values, well exceeding 1000 observation threshold after which the aforementioned issue is unlikely to apply. Moreover, the results are robust to the underlying vector error correction model specification further alleviating the concerns.

The last property that is not unique to the text but is of utmost importance is a high likelihood of extreme outliers. Unstructured data is significantly affected by processing errors. Text manipulation, such as optical character recognition (OCR) or document section identification, are imprecise. For example, the minimum of positive and negative words in an article is zero (Table 4), unlikely given the breadth of Loughran & McDonald (2011) dictionaries which include both financial and non-financial terms. Similarly, the minimum length of an article is just 36 words, potentially indicating that only a header was processed. The maximum length of a supposedly brief section is 111,162 words, more suitable for a page or an entire newspaper.

Taken altogether, these results suggest a set of requirements that need to be taken into account when transitioning from the individual word counts to a stationary measure. The procedure needs to be robust to outliers, implement a cointegration adjustment, and yield a stationary variable. Ideally, it would also make use of data's unique properties. When working with the text, these properties might include power law distribution of the word frequencies, stopword irrelevance, and context dependence.

3 Constructing a text-based measure

The main goal is to provide an alternative to simple frequency measure (a ratio of word count to text length) while accounting for the specific properties of text. There are three main problems with the frequency. First, it does not guarantee the stationarity of resulting measure. Second, the division does not account for non-stationarity of the individual word counts or text lengths.

The third issue is more prominent in a cross-sectional setting. The frequency measure might inherit statistical properties of the inverse text length. It would happen when the variation of text lengths within the sample greatly exceeds that of relevant terms. Consider an extreme example where the numerator is constant. After the division, all variation between the individual observations would come from the denominator. It is antithetical to the intent behind the division by text length which is just scaling. Moreover, the text length is correlated with financial characteristics (such as firm size and complexity, see Loughran & McDonald (2014)) resulting in an artificial introduction of bias. However, this issue is sometimes easy to diagnose. It is sufficient to replace the numerator with a constant and check if the results change. If they do not, then the issue is likely present.

Even though the news-based sentiment is of interest here, the procedure is generally applicable, can be implemented for any topic, and adapted for cross-sectional applications (for example, by avoiding differencing). The first step accounts for the outliers arising from the data processing errors. Top 1% of the individual component observations are winsorized.[9] Then the Engle-Granger two-step method is used to create stationary variables. In this context, it entails regressing relevant word counts on the text length and then differencing the residuals. The procedure can be used in-sample (future sentiment levels affect current residuals) or out-of-sample (incremental re-estimation).

[9] All subsequent results are robust. The outcome is not affected by keeping outliers as-is, varying winsorization level (1%, 5%, 10%), or using trimming instead.

$$
\begin{aligned}
log(W_{Pos,t}) &= \beta_{Pos} log(L_t) + u_t;\ e_{Pos,t} = log(W_{Pos,t}) - \hat{\beta}_{Pos} log(L_t) \\
log(W_{Neg,t}) &= \beta_{Neg} log(L_t) + u_t;\ e_{Neg,t} = log(W_{Neg,t}) - \hat{\beta}_{Neg} log(L_t) \\
\Delta Pos_t &= e_{Pos,t} - e_{Pos,t-1} \\
\Delta Neg_t &= e_{Neg,t} - e_{Neg,t-1} \\
\Delta Pess_t &= \Delta Neg_t - \Delta Pos_t
\end{aligned}
$$

L is winsorized text length, $W_{Pos(Neg)}$ are winsorized counts of positive (negative) words, t is time index. Log transform serves a dual purpose. It reduces the impact of outliers and accounts for the power law distribution of word frequencies. The change in sentiment measures are ΔPos_t, ΔNeg_t, and $\Delta Pess_t$, scaled to have a unit variance. The out-of-sample measure is constructed incrementally by including data only up to t and repeating the procedure.

The measures are stationary under both ADF and KPSS tests (Table 5). The statioanrity can be confirmed visually (Figure 2). Generally, these measures represent a change (due to differencing) in relative (sum of the regression residuals is zero) sentiment. When aggregated yearly, the top three pessimism peaks are in 1974, 1945, and 2002. They correspond to a 1973–1975 recession (along with the 1973 oil crisis), World War II, and a post dot-com bubble crash.

The Engle-Granger two-step method has a clear interpretation when applied to text. The count of relevant terms (W) can be viewed as a function of importance (I) and author-dependent attributes (A), $W = f(I, A)$. The importance is not observable unlike some of the stylistic characteristics.[10] Assuming linear relationship (or approximation) and treating text length as a proxy for style, residuals (e) are then a measure of relevant word importance. Regressing away the text length is also an adjustment "for impact across the entire collection" (Loughran & McDonald (2011)), a viable alternative to the division in a cross-sectional setting. Linear regression is an orthogonal projection so the procedure just removes the variation common to both individual word counts and text lengths. This method can be further improved by

[10] See Loughran & McDonald (2014) for a discussion of style as it pertains to financial document readability.

adding (to text length) more proxies for the idiosyncratic attributes. Count of unique words, percentage of stopwords, frequency of negation, etc. can all be easily included in the specification without affecting statistical properties and computational cost of the measure construction.

Low computational overhead is especially important in a big data setting and can potentially expand the number of applications. Simple variable construction algorithms add value for the applications relying on vast quantities of data. Methods such as Engle & Granger (1987) procedure are computationally inexpensive and have known, well-defined properties. As a result, they can be incorporated as part of a system that automatically pre-tests data (the KPSS test is also not demanding computationally) and adjusts it as necessary in real time. Alternatively, these procedures would allow to process more text or do so quicker, a desired property for both market makers and high-frequency traders (Ait-Sahalia & Saglam (2023), Pagnotta & Philippon (2018)). Kalamara et al. (2022) findings support this view. The find that a "simple count of the word uncertainty did almost as well as the more complex Boolean methods". The implications are twofold. First, the applicability of unstructured data is limited by the informational content of text itself. Second, the trade-off between method's algorithmic computational complexity and the amount of data processed may be skewed towards the quantity without losing much.

4 Dependent variables

4.1 Daily stock returns

While it is important to construct a stationary independent variable, all other inputs need to be considered as well. However, both dependent and independent variables need to be stationary to realize the benefits, mainly statistical significance of the regression coefficient of interest. In this case, structural breaks are a particular concern, even more so given that the data is daily. Rapach (2006) finds "evidence of struc-

tural instability" in the quarterly U.S. stock returns. Much higher frequency daily data is even more problematic. The underlying cause of potential breaks is especially pertinent in the context of news-based sentiment. There is evidence of negative information being deliberately released on Fridays (Doyle & Magilke (2009), Michaely et al. (2016), Rawson et al. (2022)), consistent with the Monday effect. Strategic timing makes it impossible to distinguish between the information-based effects and those due to the change in sentiment.

Strict stationarity requires immutability of the unconditional joint probability distribution with respect to time. Trading discontinuity is one potential issue; the Monday effect is a well-documented example. The market is closed on Sundays, so Mondays are not immediately preceded by a trading day. On Mondays, the mean Dow[11] return is -8.0 bps, compared to 2.4-6.0 bps on all other days (Table 6). Perhaps most telling, the absolute lowest daily return, -25.6%, has been achieved on a Monday. The lowest non-Monday return is -12.5%. Standard deviation of returns is also different; it is much higher on Mondays at 139 bps. Pairwise similarity tests confirm the disparity (Table 7). Based on both Kolmogorov-Smirnov and Anderson-Darling tests, the empirical distribution of Monday returns is different from those of all other days (Table 7, Panels A and B). The distributions of returns on Tuesdays, Wednesdays, and Thursdays are statistically identical with a mixed evidence for Fridays. More generally, the distributions on days with and without prior trading days (labelled "Breaks") are not identical (Table 7, Panel C).

4.2 Trading volume

Trading volume is another dependent variable of interest. I aggregate individual security data from CRSP[12] and then adapt Campbell et al. (1993) procedure. So, $\Delta log(Vlm)_t = log(S_t) - log(S_{t-1})$; $S_t = \sum_{\forall j} s_{t,j}$, $s_{t,j}$ is a number of shares traded in

[11] Daily Dow returns are from Historical SPDJI (daily) database, accessed through WRDS.
[12] CRSP data starts in 1926; only common shares ("SHRCD" begins with 1) are included.

stock j on day t. Similarly to stock returns, this measure can be viewed as a percent change in the aggregate trading volume. $\Delta log(Vlm)_t$ is stationary, both visually and according to the formal tests (Internet Appendix IA1). Aggregating the individual security level data has some advantages over the exchange-level statistics. The data is widely available and covers all exchanges thus avoiding the listing bias (for example, technology and life sciences companies tend to list on the NASDAQ).

5 Predictive model

5.1 Model Specification

The model specification is based on Tetlock (2007) and Garcia (2013) but restricted to one lag[13] to account for the structural breaks. There are two separate dependent variables of interest: stock returns (R_t), and a change in trading volume ($\Delta log(Vlm)_t$). M_{t-1} is a lagged sentiment measure.

$$R_t = \beta_M M_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$$
$$\Delta log(Vlm)_t = \beta_M M_{t-1} + \beta_{Vlm} \Delta log(Vlm)_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$$

Importantly, one lag is sufficient to capture most of the properties of both stock returns (Starica & Granger (2005)) and trading volume (Campbell et al. (1993)). Including more than one lag would result in structural breaks (weekends) in the controls. For example, suppose there are two lags. The breaks would just "migrate" to the independent variables, specifically R_{t-1} and R_{t-2} terms. When R_t is a Tuesday, two lags would contain Monday and Friday, in turn encapsulating the undesirable break. The same argument applies to a higher (than two) number of lags. A limitation of one lag specification is inability to evaluate reversal effects.

[13] An Internet Appendix for Garcia (2013) also includes a one lag model with results identical to the five lag version.

5.2 Inference

Assuming stationarity requirement is met, economic importance can be assessed based on magnitude and significance of the regression coefficient. However, the predictive regression uses imputed measures generated using the two-step procedure. Multiple stages increase the estimation uncertainty and affect the standard errors (Murphy & Topel (1985)). Additionally, the correlation structure of sentiment measures is not well-known. Asymptotic validity of the consistent variance estimators depends on averaging over infinitely-long series of uncorrelated features such as disturbances or clusters. These assumptions are unlikely to be satisfied by the measures constructed from unstructured data.

A solution has been proposed in Ibragimov & Muller (2010). There exists a significance level ($\alpha \leq 0.083$) such that the tail of Student's t-distribution is heavier than that of a normal distribution by more than data or model imperfections. Alternatively stated, t-distribution with a sufficiently low number of degrees of freedom is so heavy-tailed that using it for inference is robust to the additional uncertainty from the multi-stage procedure.

The Ibragimov & Muller (2010) method relies on partitioning the data and fitting the model to a small number (q) of groups, ideally each retaining statistical properties of the full sample. This yields $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_j), j = 1, ..., q$. H_{Null}: $\beta = \beta_0$ is then rejected in favor of H_{Alt}: $\beta \neq \beta_0$ if $|t_{IM}| > T^{-1}((1 - \alpha/2), q - 1)$. $T^{-1}(p, df)$ is the inverse cumulative density function of the Student's t-distribution.

$$t_{IM} = \sqrt{q}\frac{\hat{\beta}_{Avg} - \beta_0}{s_{\hat{\beta}}}$$

$$\hat{\beta}_{Avg} = q^{-1}\sum_{j=1}^{q}\hat{\beta}_j$$

$$s^2_{\hat{\beta}} = (q-1)^{-1}\sum_{j=1}^{q}(\hat{\beta}_j - \hat{\beta}_{Avg})^2$$

The main difficulty comes from partitioning the sample. Individual groups should be representative of the entire sample. The associated estimates ($\hat{\beta}_j$) should also be asymptotically independent. If these two conditions are satisfied, then the statistic itself is asymptotically valid (Ibragimov & Muller (2010)).

A specific case of long, daily time series of returns fits both conditions when partitioned as follows. Consider categorizing observations by year based on divisibility by 4. Group 1 then contains the observations where $year \equiv 0 \ mod \ 4$; Group 2 has $year \equiv 1 \ mod \ 4$, etc. Statistical properties of the entire sample are preserved within each group as they span the same year range. To the extent possible, the partitioning also makes group estimates ($\hat{\beta}_j$) asymptotically independent. Additionally, statistical characteristics of the daily returns and trading volume are retained. The individual daily observations are included in continuous, uninterrupted year-long blocks. Economically, the scheme also keeps the business cycle dating; within a year, the continuity of recessions and expansions is maintained.

6 Results

6.1 Economic implications of non-stationarity

Are sentiment-based daily stock market predictions better in expansions or recessions? Existing literature does not offer an aligned view; Garcia (2013) finds that "the predictability of stock returns using news' content is concentrated in recessions". Notably, Garcia (2013) uses non-stationary predictors and makes the conclusion based on relative magnitudes of the regression coefficients. Employing intraday data, Sun et al. (2016) obtain the opposite result, "predictability is weaker during economic recessions and largely dissipates during low trading volume days". Moreover, in a fairly restricted setting (effect of the FOMC decisions), Gardner et al. (2021) find that "news has a bigger (smaller) effect on equity prices during bad (good) times".

In general, it is dangerous to draw conclusions about relative predictability based

on the beta magnitudes. Apart from non-stationarity, there are scaling issues, perhaps best-demonstrated through a simple example. Suppose, in an otherwise identical setting, the first predictor has a beta of 3, and the second one has a beta of 6. Without affecting any predictive properties, one can multiply the second predictor by 2 thus causing the beta to fall to 3. A common rejoinder to this argument is that the predictors are often standardized to have a unit variance. However, standardization destroys informational content that is contained in the magnitude of the variance. Alternatively stated, after standardization, even though the variances are the same, the probability of observing a large swing differs. From this perspective, robust methods provide more economically valuable conclusions since they also require less variable transformations.

Table 8 includes the results obtained by predicting daily Dow returns with the non-stationary measures of news-based sentiment, constructed exactly as in Garcia (2013). Garcia (2013) measures are just word frequencies.[14] The original conclusions in Garcia (2013) are derived from the relative magnitudes of the prediction betas. All results are computed three times (all days, all days excluding breaks, breaks only) to demonstrate the ramifications of the structural breaks. The predictive specification,[15] $R_t = \beta_G G_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$, differs from Garcia (2013) in the number of lags. However, Garcia (2013) notes that the "choice of lags and controls in specification does not affect any of the conclusions of the paper", further confirmed by comparing the results.

First, as in Garcia (2013), consider all days (Table 8, Panel B, "All Dates") without any exclusions. The one-lag model retains all key results. The magnitude of β_G is larger (e.g., for the pessimism, β_G is -10.6 bps in recessions and -4.5 bps in expansions) in recessions for all measures (Table 8, Panel B). They are also identical for the positive and negative word frequencies (Table 8, Panel B) in both recessions

[14] Pos/Len, Neg/Len, and $(Neg - Pos)/Len$ (positive sentiment, negative, pessimism), all scaled to have zero mean and unit variance. Garcia (2013) averages word counts and text lengths over the period without trading to account for the weekends and other breaks.

[15] Recessions indicator is USRECD from NBER.

and expansions (3.3 and 3.5 bps in expansions, 8.6 and 8.0 bps in recessions). However, there is already a warning sign; adjusted R^2 is higher in expansions for all measures (0.66% vs 0.40% for pessimism). It is unlikely for the predictability to be concentrated in recessions and have a lower adjusted R^2 at the same time.

Excluding breaks (Table 8, Panel B, "No Breaks") further muddles the results. Generally, they are similar to the above. However, the regression coefficient magnitudes are no longer identical for the positive and negative word frequencies in recessions (positive $\beta_G = 6.7$, negative $\beta_G = -3.7$). It is not surprising that excluding breaks affects the recessions more; spurious relationships are more pronounced in small samples (Ferson et al. (2003)). Focusing on breaks (Table 8, Panel B, "Breaks") yields even more insight. There are two striking facts. First, the adjusted R^2 values go up by an order of magnitude regardless of the measure, a hallmark of spurious relationships. Second, the relationship between R^2 values in recessions and expansions flips. For the breaks, R^2 values are universally higher in recessions.

Ibragimov & Muller (2010) t-statistics provide some clarity. Despite the non-stationarity of word frequencies, none of the t_{IM} statistics are significant (Table 8, Panel B) in recessions. It is not a statistical aberration. The finding holds regardless of a predictor variable and whether the breaks are excluded or not. Consider the combination of negative word frequency and all dates without exclusions. Then the t_{IM} is -2.157 (Table 8, Panel B, "All Dates"). The associated $\hat{\boldsymbol{\beta}}$ is $(-0.000149; -0.000415; -0.001550; -0.002422)$. One of the $\hat{\boldsymbol{\beta}}$ components, -0.000415, is much larger than the others. There are only three degrees of freedom so one "partitioning outlier" makes the t_{IM} statistic insignificant.

In general, partitioning results in a small number of samples each having less observations. Reducing the number of observations within each sample accentuates spurious relationships. Meanwhile, Student's t-distribution with a low number of degrees of freedom does not allow for any of the $\hat{\boldsymbol{\beta}}$ components to significantly deviate from the mean while retaining the statistical significance. As a result, there

are two scenarios where t_{IM} is statistically significant. First, there isn't a spurious relationship and all components of the $\hat{\boldsymbol{\beta}}$ are sufficiently similar. Second, the effect of non-stationarity is uniform and the components of $\hat{\boldsymbol{\beta}}$ are affected comparably. The second scenario is rare but still feasible. In fact, it happens in expansions with the positive word frequency as a predictor (Table 8, Panel B, "All Dates"). The t_{IM} is highly significant, standing at 13.7, but it is also associated with the lowest adjusted R^2 across all measures. As a result, it is likely spurious.

Taken altogether, the regression results obtained with the Garcia (2013) measures are internally inconsistent. One way to check their validity is to use a robust, residual-based statistic while keeping the non-stationary frequency measures. The comparison between prediction errors in expansions and recessions is in Table 8, Panel A. The forecast errors are statistically significantly lower[16] in expansions, regardless of the partitioning, a reversal of the original Garcia (2013) conclusion.

Finally, non-stationarity provides an alternative explanation for the high R^2 values and large test statistics observed across the applications of text-based sentiment. Zhou (2018) notes that "in comparison with market- and survey-based measures, it is surprising that measures based on textual analysis perform better by far". His explanation for the phenomenon is based on market inefficiency, claiming that "stock market is likely to overlook information". On the contrary, supposedly "better" performance is due to a spurious relationship. Inefficiency is also not set in stone. Equity prices are an imperfect gauge of information absorption. Price signals may cancel out under the heterogeneous beliefs.

There are three main takeaways. The economic interpretation is affected by the non-stationary measures and structural breaks. Sentiment-based daily Dow return predictability is higher in expansions. Robust inference is a necessity; it either yields a correct conclusion or is very likely to be statistically insignificant.

[16]See Internet Appendix IA2 for the prediction errors obtained with the stationary two-step sentiment measures. The results are the same, forecast errors are lower in expansions.

6.2 Financial markets and news-based sentiment

Table 9, Panel C validates the out-of-sample predictive power of the stationary two-step news-based sentiment measures. The out-of sample procedure follows Welch & Goyal (2008). There is only one predictor (the measure of interest), performance is evaluated using $R^2_{OOS} = 1 - MSE_{Sent.}/MSE_{Hist.Avg.}$, and a training period is set to approximately 20% of the entire sample. All R^2_{OOS} are greater than zero so the stationary sentiment measures predict daily Dow returns and trading volume.

Table 9, Panels A and B iniclude the in-sample results. Over 1926-2005, the Dow is predicted with an in-sample adjusted R^2 of 0.09% to 0.19% depending on the sentiment measure (Table 9, Panel A). Out-of-sample, the range is 0.06% to 0.15% (Table 9, Panel C). Trading volume is also forecastable; in-sample R^2 values are 0.07% to 0.69% (Table 9, Panel B), out-of-sample 0.06% to 0.68% (Table 9, Panel C). Conditional on the 1926-2005 time period,[17] the highest in-sample adjusted R^2 for the sentiment-based Dow predictability is 0.19%; for the trading volume, it is 0.52% (Table 9, Panels A and B). In this setting, there is only one independent variable so the R^2 values may be compared directly across specifications. As a result, in-sample look-ahead bias is immaterial since both R^2 values and prediction errors are sized similarly to the out-of-sample counterparts.

Table 10 provides results for the full predictive models with controls. The control variables differ across the dependent variables and time periods, so the models are not directly comparable. Even still, the predictability pattern stays the same. Using the news-based sentiment, trading volume is easier to explain than the equity returns. When predicting the daily Dow, the change in pessimism maximizes the adjusted R^2, at 0.27%. The change in negativity does the same for the trading volume, with the adjusted R^2 getting to 2.5%. Coefficient estimates tell the same story. A one standard deviation change in pessimism elicits the largest[18] shift in market returns,

[17] Longest overlap between the equity returns and trading volume data.

[18] The measures capture predictability on the same days but some are more precise. See Internet Appendix IA4.

at 3.85 basis points (Table 10). Increasing the measure of negative sentiment by one standard deviation leads to a 1.5 percentage point decrease in transactions (Table 10).

On the surface, 3.85 bps appear economically insignificant, especially considering the historical cost and risks of implementing a corresponding trading strategy. Harnessing the premium would require calling a broker, paying a fee, waiting to place an order, taking on front running risk (the news are not exclusive - someone else might have done the same faster), counter-party risk, etc. However, there are also quiet, low market movement days. For example, the tenth percentile of absolute returns is 9 bps (Table 6). On a relative basis, 3.85 bps is a large component (3.85/9=0.423) indicating that the sentiment matters when the market movements are subdued. It is possible that the repricing of fundamentals causes the outsized market movements and the news-based sentiment is responsible for the small jitters.

Meanwhile, 1.5 percentage point decrease in the trading volume is economically tangible. For reference, the mean daily change in trading volume is 1.52% with a standard deviation of 20.8%. 1.5 percentage points can then be viewed as a difference between normal trading day and the market grinding to a halt. Granted, changes in the trading volume are hugely variable, as indicated by the standard deviation, but those are also unlikely to be driven by the sentiment. Investors need to readjust portfolios when their expectations change or the economic fundamentals get repriced. From this perspective, just like with the market returns, the relationship between trading volume and news-based sentiment is more prominent absent any shocks. Therefore, M. Baker & Wurgler (2007) suggestion to use trading volume as a proxy for sentiment can be improved. The trading volume needs to be conditioned on a low market movement.

Trading volume predictability furthers the disconnect between the news-based sentiment and market fndamentals. The trading volume is minimized with an increase in negative sentiment (Table 10, the coefficient on change in negative sentiment is less

than zero and statistically significant). First, there is less money available to use for the discretionary trading. Second, during an economic distress, the fundamentals are depressed with a very limited uncertainty about their state. Animal spirits are secondary to the rational valuation, which is responsible for the majority of substantial price changes and trading volume accumulation. Finally, bad times also involve more external interventions such as the interest rate adjustments. All these frequent shifts in the fundamentals are independent from the news and are not predictable. This view is consistent with Mai et al. (2022), who find that "the majority of assets, trade-based sentiment measures outperform their text-based equivalents for both in-sample and out-of-sample predictions". Actions are more directly linked to the economy, while news may or may not precipitate any money movement to or from the financial market.

Quantile regression provides additional evidence against the connection between extreme market movements and the news-based sentiment. It serves a dual purpose: to validate the OLS estimates (and their significance), and to check the behavior at the tails. It does not have a closed form solution for the standard errors, is less likely to be affected by the additional uncertainty from two-stage predictors, and is robust to outliers. As expected, given the large sample size, the estimates (sign, magnitude, and significance) of the conditional median ($\tau = 0.5$) closely mirror the OLS (Table 10). For the daily Dow returns and the change in pessimism, the OLS regression coefficient is -3.85 and the quantile regression is -3.98. Similarly, t_{IM} statistics for the OLS and quantile regression p-values[19] are in accord.

Expanding the range of estimated quantiles ($\tau \in [0.1, 0.9]$, 0.1 increment) demonstrates the stability of betas. The estimates[20] are relatively flat (Figure 3) and do not change with the quantile, regardless of whether the dependent variable is the daily Dow returns or the change in trading volume. The finding is an update to Tetlock

[19]Since there are no closed form solutions, statistical significance for the quantile regression is calculated using three different methods.

[20]See Internet Appendix IA3 for additional details.

(2007), who found that the "effect of negative sentiment on the Dow appears to be strongest near the extreme values of returns and sentiment". Importantly, stable, flat betas are consistent with the proposed relationship between the news-based sentiment and market gyrations. Had the sentiment played a prominent role in the extreme returns, the betas would have been tangibly different near the ends. Consequently, relatively stable quantile betas provide further statistical evidence supporting the importance of news on the quiet days with limited market movement.

Investor disagreement may potentially explain the aforementioned behavior. Heterogeneous actions are able to "generate a lot of trading activity even when prices are not moving relative to fundamentals" (Hong & Stein (2007)). Opinion-based transactions cancel each other out leading to a modest price change, if any at all. Unlike prices, trading volume accumulates (instead of cancelling out) the opposing views, so it is more sensitive and easier to predict than the returns. A negative relationship between the sentiment and trading volume is also consistent with the disagreement. As opinions get more aligned, there is less need to transact. At a firm-level, Chang et al. (2022) demonstrates that the investor disagreement has an empirically tangible effect. Taken altogether, these results show that the effects aggregate. The disagreement may be responsible for the market-wide trading patterns.

7 Conclusion

Machine learning methods receive a lot of criticism for being opaque. The issue stretches further; data that goes into these algorithms has unknown statistical properties. From this perspective, unstructured data itself is as much a "black box" as the advanced algorithms used to process it.

Among many others, there are implications for the model risk management. The data itself might lead to nonsensical results. Consequently, it is important to know the assumptions underpinning employed algorithms and check if the inputs conform.

Here, the focus is on stationarity since both complex and simple procedures alike rely on it for validity.

It is difficult to conclusively determine if the variables constructed from text are stationary or not. The KPSS test rejects trend stationarity, while, at the same time, the ADF test rejects the null of unit root presence. Visual evidence supports the KPSS test. This effect is very pronounced at a daily frequency, precisely where the alternative data adds the most value. Time series of word counts and text lengths appear remarkably similar. If a word characterization is sufficiently broad, then it should be expected for the relevant word count to be higher for a longer text. More formally, Johansen (1991) test provides strong statistical evidence supporting a cointegrating relationship. As a result, Engle & Granger (1987) procedure is a better alternative to a simple ratio of word count to text length.

There are two ways to account for non-stationarity. The easiest, although not always applicable, is to use robust statistics based on residuals (such as mean absolute error). They are sufficient to show that when market returns are forecast with the news-based sentiment, the prediction errors are higher in recessions. These results are contrary to Garcia (2013) conclusions, which are derived from comparing regression coefficients on non-stationary variables. Alternatively, it is possible to use the Engle & Granger (1987) procedure to create a stationary measure and then make economic conclusions based on the relevant coefficient's magnitude. In the same setting (predicting market returns with the news-based sentiment), a stationary measure yields an economically insignificant forecast. On the other hand, a relationship between the trading volume and news-based sentiment is much stronger.

One potential explanation for these empirical findings is that the news capture non-fundamental information. Suppose the news-based sentiment is not linked to the economic indicators but reflects the animal spirits instead. The news either reflect or influence the views of investors, subsequently causing them to act. Recessions are defined retrospectively, based on a set of objective gauges such as inflation, GDP, etc.

Then, relative to a normal expansionary state, prediction errors would also include an unaccounted component coming from the depressed fundamentals. Additionally, there is more money available for investment in expansions, so it is more likely that the news-induced urges translate to actions. Similarly, the relationship between news and trading volume is explained through the investor behavior. Given belief heterogeneity (or differences in interpretation), price signals may cancel each other but trading volume is never negative and increases with each transaction, some of which are driven by the sentiment.

Finally, unstructured data can also be of higher quality than conventional, even when economic fundamentals are of interest. One prominent example includes situations where there are trustworthiness concerns. Authoritarian countries are known to manipulate[21] or outright withhold[22] information leaving interested parties no choice but to use the alternatives. In the absence of multiple sources, it is especially important for policymakers to "know the data" since the combined forecast would be derived exclusively from the unstructured sources.

[21] The Washington Post; May 15, 2018.
[22] The Wall Street Journal; April 23, 2022.

Figures

Figure 1: **Yearly Means of Word Counts and Frequencies**

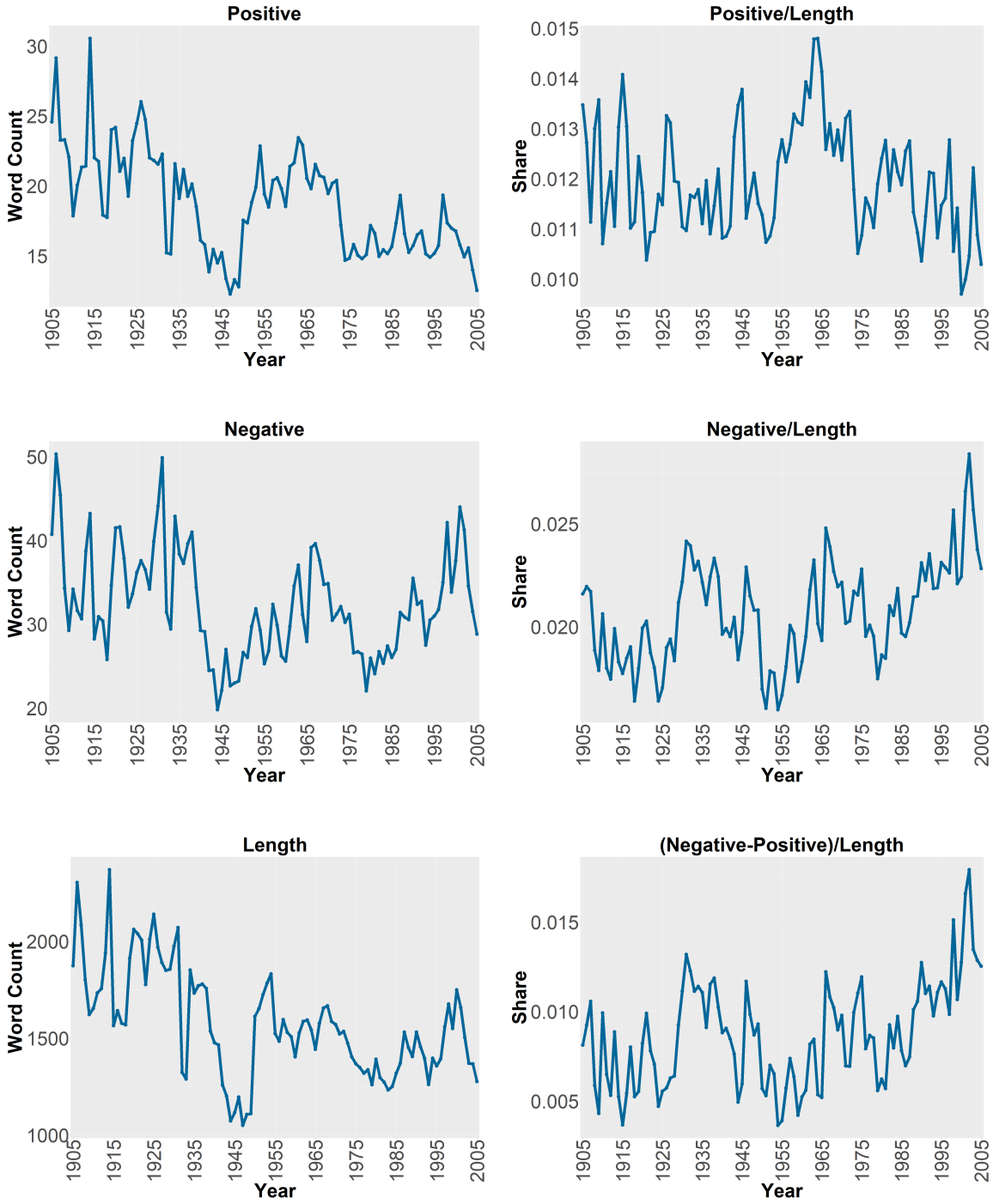


Figure 2: **Yearly Means of Stationary Change in Sentiment Measures**

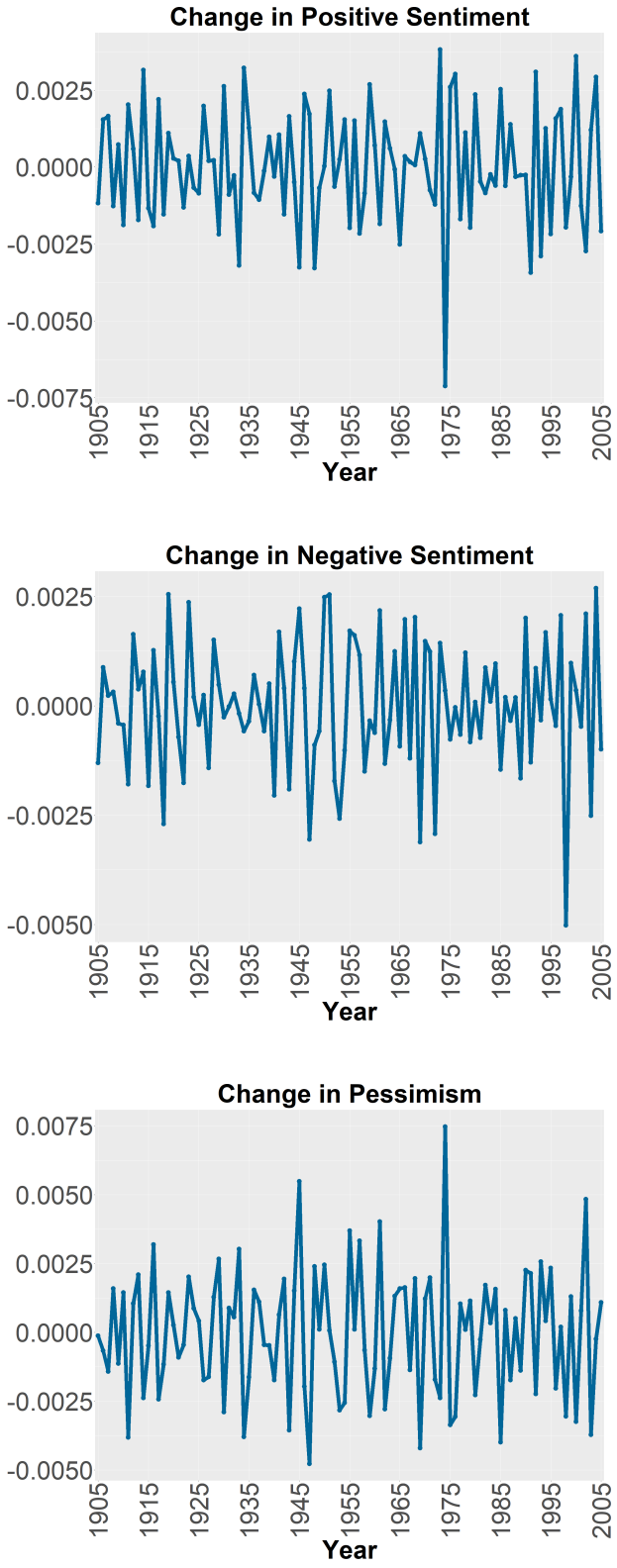
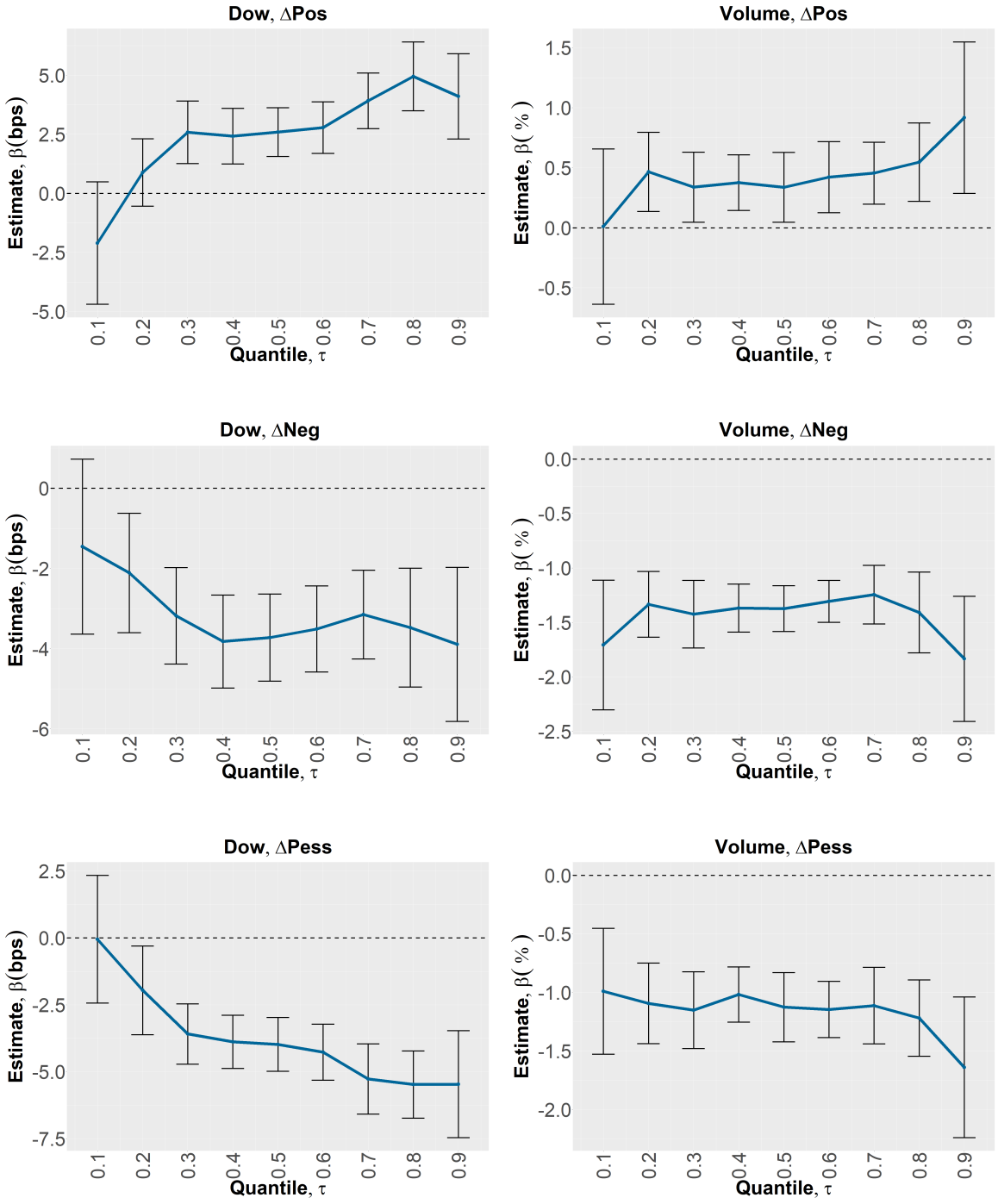


Figure 3: **Stationary Sentiment, Quantile Regressions, $\tau \in [0.1; 0.9]$**



Tables

Table 1: **Stationarity of Time Series Constructed from Unstructured Data**

Article	**Variable**	**Method**	**Stationarity**
Sentiment during Recessions, Garcia (2013)	Sentiment	Dictionary, Frequency	ADF, KPSS: reject null. Results holds for all measures and underlying components.
A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news, Obaid & Pukthuanthong (2021)	Sentiment	Multiple Neural Networks	ADF, KPSS: reject null. Results holds for all measures.
Measuring news sentiment, Shapiro et al. (2020)	Sentiment	Dictionary, SVR	ADF, KPSS: reject null.
Measuring Economic Policy Uncertainty, S. R. Baker et al. (2016)	Uncertainty	Dictionary, Frequency	ADF: reject null at 1% with some exceptions (2 states, 2 countries). KPSS (state level data): at 1%, reject null for 77/153 series; at 5%, reject null in 106 cases; at 10% in 123. KPSS (country): reject null in 17/25 cases.
News implied volatility and disaster concerns, Manela & Moreira (2017)	Volatility	Dictionary, Frequency, SVR	ADF, KPSS: reject null. Exceptions: ADF test fails to reject null hypothesis at 1% level. Natural disasters time series fails to have the null rejected under the KPSS test.
The Structure of Economic News, Bybee et al. (2020)	Multiple Topics	Latent Dirichlet Allocation	ADF: at 1%, null hypothesis is rejected 168/180 times. KPSS: at 1%, null hypothesis is rejected for 160/180 topics.
The economic effects of trade policy uncertainty, Caldara et al. (2020)	Uncertainty	Dictionary, Frequency	ADF: reject null for all text-based time series except total number of articles (monthly). KPSS: reject null.
Measuring Geopolitical Risk, Caldara & Iacoviello (2022)	Geopolitical Risk	Dictionary, Frequency	ADF: reject null for 108 series (monthly), all (daily). KPSS: reject null for 81/110 series (monthly); reject null for 6/6 (daily).

Table 2: **Stationarity Tests, Word Counts and Frequencies**

The ADF test includes intercept and trend, the number of lags is set to 1. The ADF null hypotheses: presence of a unit root (τ_3), unit root without trend (ϕ_3), unit root without trend and without drift (ϕ_2). The KPSS null hypothesis is trend-stationarity. The number of lags (KPSS) is set to $4(T/100)^{0.25}$. Critical values (cval) are at 1% level.

	ADF						**KPSS**	
	$\boldsymbol{\tau_3}$	**cval**	$\boldsymbol{\phi_2}$	**cval**	$\boldsymbol{\phi_3}$	**cval**	$\boldsymbol{\tau}$	**cval**
Length	-107.39	-3.96	3,844.51	6.09	5,766.77	8.27	3.73	0.22
Count Pos.	-104.33	-3.96	3,628.53	6.09	5,442.80	8.27	2.25	0.22
Count Neg.	-96.94	-3.96	3,132.27	6.09	4,698.41	8.27	4.19	0.22
Freq. Pos.	-97.20	-3.96	3,149.00	6.09	4,723.50	8.27	2.55	0.22
Freq. Neg.	-89.08	-3.96	2,645.13	6.09	3,967.69	8.27	2.74	0.22
Pessimism	-90.49	-3.96	2,729.67	6.09	4,094.51	8.27	3.43	0.22
Dow Ret.	-112.47	-3.96	4,216.49	6.09	6,324.73	8.27	0.03	0.22

Note: Market (Dow) returns are included for reference, a known stationary variable passes both tests.

Table 3: **Cointegration**

The results include trace and and eigenvalue-based Johansen (1991) cointegration tests. Pos, Neg, Len are non-stationary words counts, r is a number of cointegrating vectors. Null hypothesis is no cointegration. Critical values are from Osterwald-Lenum (1992).

				Critical Values		
Variables	**Type**	**Hypothesis**	**Test Stat.**	**10%**	**5%**	**1%**
Pos, Len	Trace	$r \leq 1$	7,691.72	10.49	12.25	16.26
Pos, Len	Trace	$r = 0$	17,340.50	22.76	25.32	30.45
Pos, Len	Eigen	$r \leq 1$	6,545.90	10.49	12.25	16.26
Pos, Len	Eigen	$r = 0$	16,275.93	22.76	25.32	30.45
Neg, Len	Trace	$r \leq 1$	6,843.33	10.49	12.25	16.26
Neg, Len	Trace	$r = 0$	16,112.80	22.76	25.32	30.45
Neg, Len	Eigen	$r \leq 1$	7,691.72	10.49	12.25	16.26
Neg, Len	Eigen	$r = 0$	9,648.78	16.85	18.96	23.65
Pos, Neg	Trace	$r \leq 1$	6,545.90	10.49	12.25	16.26
Pos, Neg	Trace	$r = 0$	9,730.03	16.85	18.96	23.65
Pos, Neg	Eigen	$r \leq 1$	6,843.33	10.49	12.25	16.26
Pos, Neg	Eigen	$r = 0$	9,269.47	16.85	18.96	23.65

Table 4: **Summary Statistics of the Word Counts, 1905-2005**

This table includes summary statistics for the individual word counts and text lengths. The word counts and texts lengths are at a daily frequency. All days are included. The first block shows pairwise Pearson correlation coefficients, n is a number of observations.

	Correlations								
	Pos	**Neg**	**Len**	**n**	**Mean**	**StDev**	**Min**	**Med**	**Max**
Positive	1	0.62	0.84	27,447	18.81	11.94	0	18	1,385
Negative	0.62	1	0.83	27,447	32.48	19.02	0	30	1,975
Length	0.84	0.83	1	27,447	1,583.14	847.48	36	1,530	111,162

Table 5: **Stationarity Tests, Change in Sentiment Measures**

ΔPos, ΔNeg, $\Delta Pess$ are changes in positive sentiment, negative sentiment, and pessimism constructed using the two-step procedure. The ADF null hypotheses: presence of a unit root (τ_3), unit root without trend (ϕ_3), unit root without trend and without drift (ϕ_2). The KPSS null hypothesis is trend-stationarity, alternative is presence of a unit root.

				Critical Values		
	$\boldsymbol{\Delta Pos}$	$\boldsymbol{\Delta Neg}$	$\boldsymbol{\Delta Pess}$	**10%**	**5%**	**1%**
ADF, $\boldsymbol{\tau_3}$	-196.81	-194.47	-193.74	-3.12	-3.41	-3.96
ADF, $\boldsymbol{\phi_2}$	12,911.171	12,605.58	12,511.15	4.03	4.68	6.09
ADF, $\boldsymbol{\phi_3}$	19,366.751	18,908.38	18,766.72	5.34	6.25	8.27
KPSS	0.00053	0.00038	0.00040	0.119	0.146	0.216

Table 6: **Daily DJIA Returns Summary Statistics**

"Break" is the average length of consecutive prior non-trading days. "No Breaks" ("Breaks") include only days when the market was open (closed) on a prior day; "Br. ex Mon" are non-Monday breaks. MAR is mean absolute return, AR10 is the first decile.

			Daily DJIA Returns (%)						
Group	**Count**	**Break**	**MAR**	**AR10**	**Mean**	**StDev**	**Min**	**Med**	**Max**
All Trading	25,253	0.46	0.742	0.091	0.020	1.13	-25.6	0.047	14.3
No Breaks	19,600	0	0.704	0.087	0.044	1.04	-12.5	0.055	13.9
Breaks	5,653	2.04	0.877	0.102	-0.063	1.39	-25.6	0.009	14.3
Br. ex Mon	754	1.90	0.919	0.103	0.048	1.45	-10.4	0.106	14.3
Monday	4,899	2.07	0.871	0.102	-0.080	1.39	-25.6	0	11.2
Tuesday	5,100	0.19	0.716	0.089	0.036	1.05	-12.5	0.042	13.9
Wednesday	5,143	0.03	0.727	0.089	0.055	1.10	-10.4	0.057	14.3
Thursday	5,064	0.02	0.694	0.086	0.024	1.03	-8.7	0.026	9.0
Friday	5,047	0.04	0.709	0.088	0.060	1.05	-8.8	0.094	8.9

Table 7: **Pairwise Tests of Return Distribution Similarity**

Panels A reports p-values from comparing means of returns (above the diagonal) or means of squared returns (below the diagonal) using Welch's t-test. Panel B reports p-values from comparing distributions of returns using Kolmogorov-Smirnov (KS) test (above the diagonal); Anderson-Darling (AD) test (below the diagonal). Panel C reports p-values from the same comparisons but only for the different partitions. Two-sample KS and AD p-values rely on asymptotic distributions. "Break" is a subset of days that had a non-trading prior day, "No Breaks" have no prior breaks in trading. "BxM" are breaks excluding Mondays. If p-value is less than 0.001 it is reported as 0.

Panel A: Parametric Tests

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**BxM**
Mon	1	0	0	0	0	0.024
Tue	0	1	0.369	0.570	0.239	0.826
Wed	0.001	0.251	1	0.144	0.800	0.897
Thu	0	0.673	0.107	1	0.080	0.664
Fri	0	0.995	0.218	0.632	1	0.820
BxM	0.669	0.006	0.015	0.004	0.005	1

Welch's t-test, Returns

Welch's t-test, Returns^2

Panel B: Non-parametric Tests

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**BxM**
Mon	1	0	0	0	0	0.012
Tue	0	1	0.753	0.118	0.006	0.020
Wed	0	0.873	1	0.332	0.019	0.007
Thu	0	0.194	0.140	1	0.001	0.001
Fri	0	0.018	0.022	0.002	1	0.013
BxM	0.006	0.001	0.001	0	0.0005	1

KS Test

AD Test

Panel C: Breaks

	Welch's t, Ret.	**Welch's t, Ret.^{2}**	**KS**	**AD**
No Breaks vs Breaks	0	0	0	0
No Breaks vs BxM	0.937	0.004	0.006	0.0001

Table 8: **Robust Hypothesis Testing, Word Frequencies**

Specification: $R_t = \beta_G G_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$. Word frequencies ($G$) are Pos/Len, Neg/Len, and $(Neg - Pos)/Len$ (positive, negative, pessimism) constructed as in Garcia (2013). Word frequencies are non-stationary. Dow returns span 1905-2005. MAE and MSE are compared using two-sample Welch's t-test; RMSE is reported instead to match units. All p-values less than 0.001 are entered as 0.

Panel A: Dow Returns, Prediction Errors

	MAE (bps)			RMSE (bps)		
	Exp	**Rec**	**p-value**	**Exp**	**Rec**	**p-value**
All Dates						
Pos. Freq.	67.205	97.386	0	97.555	153.343	0
Neg. Freq.	67.186	97.233	0	97.548	153.362	0
Pessimism	67.186	97.146	0	97.514	153.250	0
No Breaks						
Pos. Freq.	63.851	91.545	0	90.295	140.592	0
Neg. Freq.	63.842	91.497	0	90.291	140.676	0
Pessimism	63.840	91.435	0	90.275	140.609	0
Breaks						
Pos. Freq.	78.285	116.702	0	118.452	187.469	0
Neg. Freq.	78.220	116.060	0	118.455	186.983	0
Pessimism	78.198	116.031	0	118.286	186.749	0

Panel B: Robust t-stat

	β_G (bps)		t_{IM}		$cv, \alpha/2$		Adj. R^2 (%)	
	Exp	**Rec**	**Exp**	**Rec**	**.005**	**.025**	**Exp**	**Rec**
All Dates								
Pos. Freq.	3.296	8.558	13.664	2.368	±5.841	±3.182	0.572	0.282
Neg. Freq.	-3.535	-7.950	-3.106	-2.157	±5.841	±3.182	0.587	0.258
Pessimism	-4.542	-10.590	-4.355	-2.330	±5.841	±3.182	0.655	0.402
No Breaks								
Pos. Freq.	2.251	6.690	4.155	1.851	±5.841	±3.182	0.482	0.123
Neg. Freq.	-2.416	-3.709	-2.395	-1.411	±5.841	±3.182	0.491	0.003
Pessimism	-3.074	-6.252	-2.812	-1.856	±5.841	±3.182	0.527	0.099
Breaks								
Pos. Freq.	7.920	16.262	4.162	2.556	±5.841	±3.182	2.033	2.738
Neg. Freq.	-8.071	-23.622	-2.321	-2.329	±5.841	±3.182	2.028	3.241
Pessimism	-10.906	-26.801	-3.995	-2.454	±5.841	±3.182	2.307	3.484

Table 9: **Out-of-Sample Validation of Predictability**

Model specifications: $R_t = \beta_M M_{t-1} + C + \epsilon$ and $\Delta log(Vlm)_t = \beta_M M_{t-1} + C + \epsilon$. Training periods are 1905-1925 (Dow), 1926-1941 (trading volume). Only days with no prior breaks are included. Sentiment measures are stationary, created with the two-step procedure. Return regression errors are in basis points (bps), trading volume in percentage points (%pp). Out-of-sample measures are incrementally re-estimated and are benchmarked against the historical average.

Panel A: In-Sample, Daily Dow Returns						
	1905-2005			**1926-2005**		
	MAE (bps)	***RMSE*** (bps)	***Adj.R²*** (%)	***MAE*** (bps)	***RMSE*** (bps)	***Adj.R²*** (%)
ΔPos	70.170	104.075	0.058	70.131	105.695	0.092
ΔNeg	70.128	104.047	0.111	70.085	105.671	0.138
ΔPess	70.123	104.035	0.135	70.065	105.646	0.186
Panel B: In-Sample, Trading Volume						
	1926-2005			**1942-2005**		
	MAE (%pp)	***RMSE*** (%pp)	***Adj.R²*** (%)	***MAE*** (%pp)	***RMSE*** (%pp)	***Adj.R²*** (%)
ΔPos	14.7843	20.7606	0.0370	12.8599	17.6960	0.0695
ΔNeg	14.7390	20.7103	0.5204	12.8133	17.6413	0.6861
ΔPess	14.7571	20.7320	0.3115	12.8305	17.6622	0.4507
Panel C: Out-of-Sample, Returns and Volume						
	Returns, 1926-2005			**Volume, 1942-2005**		
	MAE (bps)	***RMSE*** (bps)	R^2_{OOS} (%)	***MAE*** (%pp)	***RMSE*** (%pp)	R^2_{OOS} (%)
Benchmark	70.1996	105.7556	0	12.8695	17.7062	0
ΔPos	70.1617	105.7244	0.0591	12.8644	17.7009	0.0598
ΔNeg	70.1235	105.6919	0.1204	12.8175	17.6460	0.6788
ΔPess	70.1117	105.6760	0.1505	12.8332	17.6664	0.4496

Table 10: **Dow and Volume Predictability, Stationary Sentiment**

Model specifications: $R_t = \beta_M M_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$ and $\Delta log(Vlm)_t = \beta_M M_{t-1} + \beta_{Vlm} \Delta log(Vlm)_{t-1} + \beta_R R_{t-1} + \beta_{RSq} R^2_{t-1} + C + \epsilon$. Only days with no prior breaks are included. Sentiment measures are stationary, created with the two-step procedure. Return regression betas are in basis points, trading volume in percentage points. Adj. R^2 is in percent, p-values less than 0.0001 are reported as 0. Quantile regression confidence intervals follow Koenker (1994), kernel p-value is based on Powell (1991), bootstrap is pairwise as described in Koenker (2005).

	Dow Returns 1905-2005			**Volume 1926-2005**		
	$\boldsymbol{\Delta Pos}$	$\boldsymbol{\Delta Neg}$	$\boldsymbol{\Delta Pess}$	$\boldsymbol{\Delta Pos}$	$\boldsymbol{\Delta Neg}$	$\boldsymbol{\Delta Pess}$
$\boldsymbol{\beta_{M,OLS}}$	2.454	-3.472	-3.853	0.376	-1.510	-1.162
$\boldsymbol{t_{IM}}$	7.199	-6.367	-11.908	1.656	-15.111	-5.484
$\boldsymbol{cv, \alpha/2 = .025}$	±3.182	±3.182	±3.182	±3.182	±3.182	±3.182
$\boldsymbol{cv, \alpha/2 = .005}$	±5.841	±5.841	±5.841	±5.841	±5.841	±5.841
Adj. $\boldsymbol{R^2}$ **(%)**	0.190	0.243	0.266	1.990	2.468	2.253
$\boldsymbol{\beta_{M,QR}; \tau = 0.5}$	2.591	-3.720	-3.979	0.337	-1.373	-1.126
Lower Bound, $\boldsymbol{\beta_{M,QR}}$	1.511	-5.042	-4.962	0.039	-1.612	-1.419
Upper Bound, $\boldsymbol{\beta_{M,QR}}$	3.572	-2.869	-2.954	0.620	-1.192	-0.827
Bootstrap, p-value	0	0	0	0.051	0	0
Kernel, p-value	0.0001	0	0	0.040	0	0

References

Ait-Sahalia, Y., & Saglam, M. (2023, 04). High frequency market making: The role of speed. *Journal of Econometrics*. doi: 10.1016/j.jeconom.2022.12.015

Baker, M., & Wurgler, J. (2002, 02). Market timing and capital structure. *The Journal of Finance*, *57*, 1-32. doi: 10.1111/1540-6261.00414

Baker, M., & Wurgler, J. (2007, 04). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*, 129-151. doi: 10.1257/jep.21.2.129

Baker, S. R., Bloom, N., & Davis, S. J. (2016, 07). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*, 1593-1636. doi: 10.1093/qje/qjw024

Barsky, R. B., & Sims, E. R. (2012, 06). Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review*, *102*, 1343-1377. doi: 10.1257/aer.102.4.1343

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2020, 01). *The structure of economic news.* Retrieved 2023-03-10, from https://www.nber.org/papers/w26648

Caldara, D., & Iacoviello, M. (2022, 04). Measuring geopolitical risk. *American Economic Review*, *112*, 1194-1225. doi: 10.1257/aer.20191823

Caldara, D., Iacoviello, M., Molligo, P., Prestipino, A., & Raffo, A. (2020, 01). The economic effects of trade policy uncertainty. *Journal of Monetary Economics*, *109*, 38–59. doi: 10.1016/j.jmoneco.2019.11.002

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993, 11). Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, *108*, 905–939. doi: 10.2307/2118454

Chang, Y.-C., Hsiao, P.-J., Ljungqvist, A., & Tseng, K. (2022, 06). Testing disagreement models. *The Journal of Finance*, *77*, 2239-2285. doi: 10.1111/jofi.13137

Dickey, D. A., & Fuller, W. A. (1979, 06). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 427-431. doi: 10.2307/2286348

Doyle, J. T., & Magilke, M. J. (2009). The timing of earnings announcements: An examination of the strategic disclosure hypothesis. *The Accounting Review*, *84*, 157–182.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*, 251-276. doi: 10.2307/1913236

Ferson, W. E., Sarkissian, S., & Simin, T. T. (2003, 07). Spurious regressions in financial economics? *The Journal of Finance*, *58*, 1393-1413. doi: 10.1111/1540-6261.00571

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*, 1267–1300.

Gardner, B., Scotti, C., & Vega, C. (2021, 11). Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. *Journal of Econometrics*. doi: 10.1016/j.jeconom.2021.07.014

Gentzkow, M., Kelly, B., & Taddy, M. (2019, 09). Text as data. *Journal of Economic Literature*, *57*, 535-574. doi: 10.1257/jel.20181020

Gonzalo, J., & Lee, T.-H. (1998, 09). Pitfalls in testing for long run relationships. *Journal of Econometrics*, *86*, 129-154. doi: 10.1016/s0304-4076(97)00111-5

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hansen, S., McMahon, M., & Tong, M. (2019, 12). The long-run information effect of central bank communication. *Journal of Monetary Economics*, *108*, 185-202. doi: 10.1016/j.jmoneco.2019.09.002

Haug, A. A. (1996, 03). Tests for cointegration a monte carlo comparison. *Journal of Econometrics*, *71*, 89-115. doi: 10.1016/0304-4076(94)01696-8

Ho, M. S., & Sorensen, B. E. (1996). Finding cointegration rank in high dimensional systems using the johansen test: An illustration using data based monte carlo simulations. *The Review of Economics and Statistics*, *78*, 726–732. doi: 10.2307/2109959

Hong, H., & Stein, J. C. (2007, 04). Disagreement and the stock market. *Journal of Economic Perspectives*, *21*, 109-128. doi: 10.1257/jep.21.2.109

Ibragimov, R., & Muller, U. K. (2010, 10). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, *28*, 453-468. doi: 10.1198/jbes.2009.08046

Johansen, S. (1991, 11). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, *59*, 1551. doi: 10.2307/2938278

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022, 06). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, *37*. doi: 10.1002/jae.2907

Koenker, R. (1994). Confidence intervals for regression quantiles. *Asymptotic Statistics*, 349-359. doi: 10.1007/978-3-642-57984-4_29

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992, 10). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159-178. doi: 10.1016/0304-4076(92)90104-y

Loughran, T., & McDonald, B. (2011, 01). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*, 35-65. doi: 10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2014, 07). Measuring readability in financial disclosures. *The Journal of Finance*, *69*, 1643-1671. doi: 10.1111/jofi.12162

Mai, D., Pukthuanthong, K., & Zhou, G. (2022). Investor sentiment and asset returns: Actions speak louder than words. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4281161

Manela, A., & Moreira, A. (2017, 01). News implied volatility and disaster concerns. *Journal of Financial Economics*, *123*, 137-162. doi: 10.1016/j.jfineco.2016.01.032

Michaely, R., Rubin, A., & Vedrashko, A. (2016, 10). Are friday announcements special? overcoming selection bias. *Journal of Financial Economics*, *122*, 65-85. doi: 10.1016/j.jfineco.2016.05.006

Murphy, K. M., & Topel, R. H. (1985, 10). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *3*, 370. doi: 10.2307/1391724

Newey, W. K., & West, K. D. (1987, 05). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703. doi: 10.2307/1913610

Obaid, K., & Pukthuanthong, K. (2021, 06). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics*, *144*. doi: 10.1016/j.jfineco.2021.06.002

Osterwald-Lenum, M. (1992, 08). A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics1. *Oxford Bulletin of Economics and Statistics*, *54*, 461-472. doi: 10.1111/j.1468-0084.1992.tb00013.x

Pagnotta, E. S., & Philippon, T. (2018). Competing on speed. *Econometrica*, *86*, 1067–1115.

Powell, J. (1991). *Nonparametric and semiparametric methods in econometrics and statistics : proceedings of the fifth international symposium in economic theory and econometrics* (W. A. Barnett, J. Powell, & G. E. Tauchen, Eds.). Cambridge University Press.

Rapach, D. E. (2006, 03). Structural breaks and predictive regression models of aggregate u.s. stock returns. *Journal of Financial Econometrics*, *4*, 238-274. doi: 10.1093/jjfinec/nbj008

Rawson, C., Twedt, B., & Watkins, J. (2022, 10). Managers' strategic use of concurrent disclosure: Evidence from 8-k filings and press releases. *The Accounting Review*. doi: 10.2308/tar-2021-0088

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020, 11). Measuring news sentiment. *Journal of Econometrics*, *228*. doi: 10.1016/j.jeconom.2020.07.053

Starica, C., & Granger, C. (2005, 08). Nonstationarities in stock returns. *Review of Economics and Statistics*, *87*, 503-522. doi: 10.1162/0034653054638274

Stock, J. H., & Watson, M. W. (2019). *Introduction to econometrics*. Pearson.

Sun, L., Najand, M., & Shen, J. (2016, 12). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, *73*, 147-164. doi: 10.1016/j.jbankfin.2016.09.010

Tetlock, P. C. (2007, 05). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139-1168. doi: 10.1111/j.1540-6261.2007.01232.x

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*, 1455–1508.

White, H. (1980, 05). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817. doi: 10.2307/1912934

Zhou, G. (2018, 11). Measuring investor sentiment. *Annual Review of Financial Economics*, *10*, 239-259. doi: 10.1146/annurev-financial-110217-022725