Statistical Properties of News Sentiment and Implications for Return Predictability*

Vitaliy Ryabinin†
Indiana University Northwest
viryab@iu.edu
www.vr19.org

Abstract

Using text-based variables with unknown statistical properties alongside conventional numeric variables carries a high risk of distorting economic inference. I find that daily news-based variables, regardless of their construction methodology, are often nonstationary and thus violate common assumptions of time series analysis. Examining news sentiment, I show that relying on existing robust methods reduces the risk of Type I error. Contrary to Garcia (2013), I find that daily news sentiment forecasts stock market returns at least as effectively during both recessions and expansions, with some evidence suggesting better predictability during expansions.

Keywords: News Sentiment, Return Predictability, Heterogeneous Data, Language Models, Inference.

JEL Classification: C13, E32, G17.

*I thank Rustam Ibragimov, Alexander Michaelides, Savitar Sundaresan, Rajkamal Iyer, Ansgar Walther, and seminar participants at Imperial College London for useful comments and suggestions. Remaining errors are my own.

†Indiana University Northwest, 3400 Broadway, Dunes Medical/Professional Building 1127, Gary, IN 46408.

1 Introduction

Measures constructed from unstructured data, such as text and images, are gaining increasing prominence, primarily due to their high frequency and relatively low cost. The upside is availability – using text expands the range of feasible empirical applications. The downside is also clear: statistical properties of unstructured data remain underexplored, and the algorithms used to process them (especially AI/ML) are often opaque. These measures are rarely the main focus; instead, they are typically used as inputs within a broader economic model. However, relying on variables with unknown properties poses a significant risk of violating the assumptions of the broader model, potentially distorting economic inference.

In this paper, I document statistical properties of news-based measures that are likely to violate the assumptions underpinning methods based on the least squares estimation. Most importantly, I find that daily text-based variables are often nonstationary. To empirically demonstrate the effect on economic inference, I reevaluate the main finding of Garcia (2013), namely that "the predictability of stock returns using news' content is concentrated in recessions." This finding is based on comparing the magnitudes of predictive regression coefficients during recessions and expansions. However, nonstationarity directly affects regression betas, leading to either unreliable or nonsensical estimates. As a result, the entire difference in the magnitudes of regression betas between recessions and expansions is explained by two factors: spurious correlation and interaction with equity market volatility. By using robust methods, I show that daily news sentiment forecasts stock market returns at least as effectively in both recessions and expansions, with some evidence suggesting better predictability during expansions.

I also examine the underlying components of the daily news sentiment measures in Garcia (2013): term counts, frequencies, and text lengths. These basic building blocks are cointegrated; term counts and text lengths share a common underlying stochastic trend. I also find that term counts, frequencies, and text lengths are nonstationary. As a result, the nonstationarity of daily news-based variables originates in the language itself and is inherited from their underlying components. Detecting nonstationarity in daily text-based variables

proves surprisingly challenging. Both the measures and their underlying components are visually nonstationary, but common unit root tests yield inconsistent results. With very few exceptions, the augmented Dickey-Fuller (ADF, Dickey and 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. This situation is rare compared to conventional numeric data. For example, in Kwiatkowski et al. (1992), out of fourteen conventional economic variables, only the industrial production series displays "evidence against both hypotheses."

From this perspective, all measure construction algorithms (e.g., term frequency, AI/ML) are likely to produce nonstationary daily news-based variables, unless they are specifically designed to handle nonstationary and often cointegrated inputs. I verify that the nonstationarity of news-based measures is a general property – rather than an artifact of the construction algorithm or data sources in Garcia (2013) – by testing several prominent daily news-based variables recently introduced in the academic literature. Specifically, I test a daily news-based sentiment measure proposed by Shapiro et al. (2022), a news-implied volatility measure (NVIX) introduced by Manela and Moreira (2017), and multiple measures of economic conditions constructed by Bybee et al. (2024). The same pattern – visual nonstationarity and inconsistent ADF and KPSS test conclusions – reemerges for 186 out of 188 daily variables. Additionally, the news-based measures of economic conditions in Bybee et al. (2024) are available at both daily and monthly frequencies. I use this feature to argue that the combination of high frequency and the type of underlying source material (text) jointly contribute to the inconsistency in unit root test conclusions.

Daily news-based measures are rarely the focus of analysis; more commonly, they serve as inputs into subsequent economic mechanism. Nonstationary inputs, however, impact a full spectrum of econometric models. For instance, common methods that incorporate regularization or variable selection, such as LASSO, Ridge, or Elastic Net, inherit the stationarity assumption from ordinary least squares estimation. Most common machine learning methods either directly assume that the data generation process remains constant over time or

estimate (update) their parameters using techniques that explicitly rely on stationarity. In fact, adapting machine learning methods to nonstationarity remains an active area of research in computer science (Sugiyama and Kawanabe (2012)). The impact of nonstationary inputs on AI algorithms remains uncertain, at least in part due to its black-box nature. In financial and economic contexts, Gentzkow et al. (2019) and Ash and Hansen (2023) provide a comprehensive overview of text processing algorithms, their underlying assumptions, intended use cases, and prominent applications.

The main risk of using text-based variables without explicitly accounting for their statistical properties is the potential to distort economic inference. There are two likely sources of distortion: biased parameters of interest or invalid confidence intervals. Importantly, incorporating any text-based variable into an economic model also introduces an imputed regressor problem (Pagan (1984), Murphy and Topel (1985)), further invalidating confidence intervals. There are only a few studies that acknowledge and attempt to address these issues. For example, to achieve reliable economic inference when model inputs are based on unstructured data, Battaglia et al. (2024) proposes using either "joint maximum likelihood estimation of the regression model and the variables of interest" or "an explicit bias correction with bias-corrected confidence intervals." Unfortunately, both of these approaches require access to the underlying source data and, as a result, greatly complicate reusing the already existing AI- or ML-generated indices.

Focusing on the relative predictability of daily equity market returns during recessions and expansions with the news-based sentiment variables, I show that robust statistics improve economic inference. I use residual-based statistics (mean absolute error, mean squared error, out-of-sample pseudo R^2, etc.), robust confidence intervals (Ibragimov and Muller (2010)), and benchmarking (adapting the methodology introduced in Welch and Goyal (2008) and Goyal et al. (2024)) to demonstrate that news-based sentiment lacks excess forecasting power in both recessions and expansions. In fact, news-based sentiment does not contain any investment information beyond what is already incorporated into prior market returns. As a result, its predictive performance is either comparable to or worse than that of basic

benchmark strategies that rely solely on historical market returns, such as the long-term and short-term averages. This relative lack of predictive performance is consistent across different time periods, observed across multiple measures, and present both in-sample and out-of-sample.

Findings in this study also complement documented economic properties of news-based sentiment. For example, by comparing R^2 values of news-based and conventional measures, Zhou (2018) notes: "In comparison with market- and survey-based measures, it is surprising that measures based on textual analysis perform better by far. This may indicate that the stock market is likely to overlook information from the latter." Instead, the observed high R^2 values are just as likely to be artifacts of nonstationarity.

Finally, employing robust inference methods to reanalyze empirical evidence can strengthen the credibility of some theoretical models. In this case, noise trading (Shleifer and Summers (1990)) and investor disagreement (Hong and Stein (2007)) models can justify the relative lack of forecasting power and resulting identical equity market return predictability during both recessions and expansions. For example, asset prices reflect noise trading when demand shifts are correlated and investors react similarly to the same noisy signal. Arbitrageurs counteracting noise trading likely operate throughout all stages of the business cycle, leading to predictability that is similar to and indistinguishable from prior return benchmarks. Investor disagreement models also similarly explain the empirical findings. Both negative and positive signals may be misinterpreted, leading to opposing trades without affecting prices. If the level of investor disagreement is independent of the business cycle, then news-based sentiment (even if it serves as a signal to all investors) would have no forecasting power. This mechanism would also be correlated with trading volume, a common market-based proxy for investor sentiment. While these mechanisms align with empirical evidence, the interaction between sentiment, the business cycle, and financial markets remains uncertain, if it exists at all. Subsequently, coupling robust inference with text-based variables represents a crucial step toward accumulating further empirical evidence, not only in the context of news-based sentiment but also across other economic and financial applications.

2 Statistical Properties of Business News

2.1 Data

This paper mainly uses the measures of positivity, negativity, and pessimism proposed in Garcia (2013) spanning 1905 to 2005. These sentiment variables are constructed from the daily counts of positive terms (*#Positive*), negative terms (*#Negative*), and text lengths (*#Length*) sourced from two New York Times columns, "Financial Markets" and "Topics in Wall Street."[1] Each term's tone has been classified as either positive or negative using the Loughran and McDonald (2011) lexicon. Then, the measures are defined as follows (t denotes the newspaper publication date).[2]

$$\begin{aligned} Positivity_t &= \#Positive_t/\#Length_t \\ Negativity_t &= \#Negative_t/\#Length_t \\ Pessimism_t &= (\#Negative_t - \#Positive_t)/\#Length_t, \end{aligned}$$

Following Garcia (2013), Dow Jones Industrial Average (Dow) log-returns are used to represent the equity market price movements. To demonstrate external validity, the results are replicated and also extended past 2005 using the San Francisco Fed daily news sentiment index (Shapiro et al. (2022)) that ranges from 1980 to 2024.

Almost any procedure used to represent raw text as a numeric array introduces noise into the resulting measurement. For example, the minimum counts of positive and negative words in this sample are both zero (Table 1), which is unlikely to be accurate given the breadth of Loughran and McDonald (2011) lexicon. Additionally, financial documents, even those as commonplace as newspapers, are often structurally complex. They may contain more than one section, with only a subset being of interest. Identifying a relevant part of the text

[1] These counts were obtained directly from Diego Garcia's website, https://leeds-faculty.colorado.edu/garcia/data.html.

[2] See Garcia (2013) for a detailed description of the measure construction procedure, which includes slight adjustments for market closures.

("Financial Markets" and "Topics in Wall Street" columns) is a difficult classification task even for an advanced AI algorithm. The maximum length of an article is 111,162 words, likely corresponding to an entire newspaper or multiple sections. The minimum length of an article in this sample is just 36 words, indicating that only a header was processed (Table 1). Finally, some of the newspapers are only available as images, requiring the use of optical character recognition, and introducing additional noise into the variable construction process.

Table 1: **Sample Statistics: Term Counts, Sentiment Measures, and Dow Returns**

This table includes summary statistics for the individual term counts and text lengths compiled in Garcia (2013). Term counts and texts lengths are daily and are grouped by the business cycle. *#Positive*, *#Negative*, *#Length* are term counts. *Positivity*, *Negativity*, and *Pessimism* are frequency measures defined as a term count divided by text length (e.g. *Positivity=#Positive/#Length*) and are reported as percentages. P25 and P75 are 25th and 75th quantiles. Recession indicator is NBER USRECD. Daily Dow log-returns are in percent.

	Obs.	**Mean**	**StDev**	**Min**	**P25**	**Median**	**P75**	**Max**
				All Dates				
#Positive	27,449	18.81	11.94	0	13	18	23	1,385
#Negative	27,449	32.48	19.02	0	22	30	40	1,975
#Length	27,449	1,583.17	847.46	36	1,301	1,530	1,747	111,162
Positivity, %	27,449	1.20	0.42	0	0.90	1.16	1.46	3.70
Negativity, %	27,449	2.06	0.67	0	1.59	1.99	2.45	6.64
Pessimism, %	27,449	0.86	0.88	-3.14	0.26	0.81	1.40	6.64
Dow, %	27,449	0.02	1.07	-25.63	-0.45	0.04	0.53	14.27
				Recessions				
#Positive	6,455	19.79	19.33	0	13	18	24	1,385
#Negative	6,455	35.75	29.74	1	24	33	43	1,975
#Length	6,455	1,717.58	1,497.64	206	1,369	1,616	1,881	111,162
Positivity, %	6,455	1.16	0.39	0	0.89	1.12	1.39	3.12
Negativity, %	6,455	2.09	0.64	0.33	1.63	2.03	2.46	5.87
Pessimism, %	6,455	0.93	0.83	-1.88	0.36	0.89	1.44	5.32
Dow, %	6,455	-0.03	1.42	-13.72	-0.62	0	0.57	14.27
				Expansions				
#Positive	20,994	18.51	8.43	0	13	18	23	99
#Negative	20,994	31.48	14.02	0	22	30	39	164
#Length	20,994	1,541.84	492.15	36	1,285	1,506	1,697	10,220
Positivity, %	20,994	1.21	0.43	0	0.90	1.17	1.48	3.70
Negativity, %	20,994	2.05	0.68	0	1.57	1.98	2.45	6.64
Pessimism, %	20,994	0.84	0.89	-3.14	0.24	0.78	1.38	6.64
Dow, %	20,994	0.04	0.94	-25.63	-0.42	0.05	0.51	9.67

2.2 Stationarity of Term Counts and Sentiment Measures

Text-based variables serve as inputs into a variety of econometric models. These econometric models either directly assume that the inputs are stationary or rely on methods requiring stationarity for parameter estimation. In this section, I demonstrate that the news sentiment measures are often nonstationary, regardless of the construction methodology (e.g., simple frequency or AI/ML). Nonstationarity of the measures is inherited from the underlying building blocks such as term counts and text lengths. I also document that it is not clear how to test the measures, term frequencies, or term counts for stationarity. Conventional unit root tests, such as the ADF and KPSS procedures, frequently yield inconsistent results. When visual indicators are used as a tiebreaker, the KPSS test appears to be more reliable than the ADF test for detecting nonstationarity in text-based variables.

Table 2 shows the results of testing term counts and term frequencies for stationarity using the ADF and KPSS tests. These tests yield inconsistent conclusions. For all term counts and frequencies, the null hypothesis is rejected under all three specifications of the ADF test: presence of a unit root, a unit root without a trend, and a unit root without a trend or drift. At face value, this suggests that there is no unit root present and that the time series are likely stationary. On the other hand, the KPSS test reaches the opposite conclusion. The null hypothesis of the KPSS test is trend stationarity, which is rejected for all term counts and sentiment measures. It is rare for the ADF and KPSS tests to produce this pattern of contradictory outcomes. For example, when the KPSS test was empirically validated using monthly and quarterly economic data, only one (industrial production) out of 14 time series displays inconsistent ADF and KPSS results (Kwiatkowski et al. (1992)).

One possible explanation for this phenomenon is suggested by Jiang et al. (2020), who observe that the KPSS test "always rejects stationarity or has no nontrivial power at high frequency." If this explanation applies here, then a conventional numeric variable observed at a similarly high frequency (daily) should exhibit similar behavior. To demonstrate that the frequency alone is insufficient to account for the inconsistent ADF and KPSS outcomes, I include daily Dow log-returns alongside term frequencies and counts (Table 2). Although

daily Dow log-returns match the frequency of the news-based variables, the ADF and KPSS tests agree. Next, to provide further evidence supporting the validity of KPSS test results, I compute yearly arithmetic means of article lengths and the corresponding yearly pessimism frequencies. To make a visual comparison and highlight the time dependency of the text characteristics, I place these means alongside the Dow log-returns, which are aggregated similarly and represent a known stationary variable. In line with the KPSS test, the article lengths and term frequencies appear non-stationary, particularly when compared to the daily Dow log-returns (Figure 1). Specifically, text-based time series do not oscillate around a stable value; their respective means are time-dependent. Using the counts of terms (*#Positive*, *#Negative*) instead of the article lengths or relying on another sentiment measure (for example, replacing *Pessimism* with either *Positivity* or *Negativity*) does not affect the findings. Overall, this behavior of the time series suggests that the type of source material (text) is a major contributor to the inconsistent ADF and KPSS outcomes.

Table 2: **Stationarity of Financial Term Counts and Sentiment Measures**

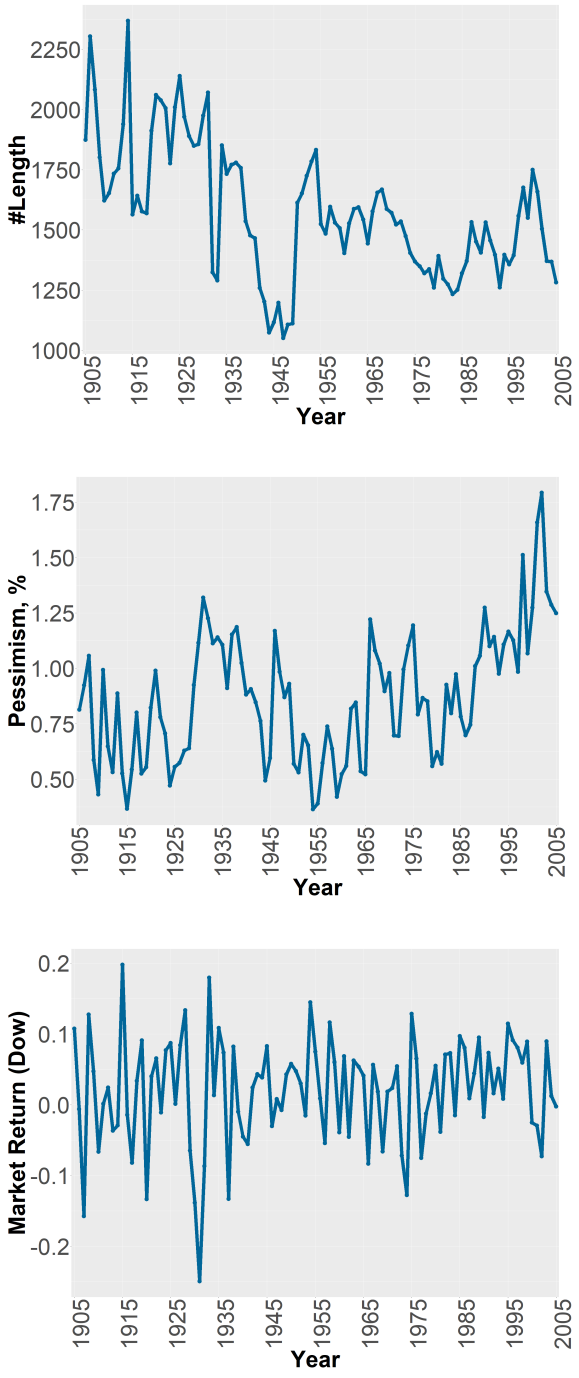
This table tests raw term counts and sentiment measures (n=27,449 days) for stationarity using ADF and KPSS tests. *#Positive*, *#Negative*, *#Length* are term counts. *Positivity*, *Negativity*, and *Pessimism* are frequency measures defined as a term count divided by text length (e.g. *Positivity=#Positive/#Length*). ADF test includes both intercept and trend, the number of lags is selected using AIC. 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 number of lags for KPSS test is set to $4(T/100)^{0.25}$; KPSS null hypothesis is trend-stationarity. Reported critical values (cval) are at 1% level. The results are robust to the lag selection.

	ADF						**KPSS**	
	$\boldsymbol{\tau_3}$	**cval**	$\boldsymbol{\phi_2}$	**cval**	$\boldsymbol{\phi_3}$	**cval**	$\boldsymbol{\tau}$	**cval**
#Positive	-104.33	-3.96	3,628.46	6.09	5,442.69	8.27	2.25	0.22
#Negative	-96.94	-3.96	3,132.46	6.09	4,698.70	8.27	4.19	0.22
#Length	-107.39	-3.96	3,844.23	6.09	5,766.35	8.27	3.73	0.22
Positivity, %	-97.21	-3.96	3,150.11	6.09	4,725.17	8.27	2.54	0.22
Negativity, %	-89.08	-3.96	2,645.28	6.09	3,967.92	8.27	2.74	0.22
Pessimism, %	-90.50	-3.96	2,730.18	6.09	4,095.27	8.27	3.43	0.22
Dow, %	-118.71	-3.96	4,697.30	6.09	7,045.95	8.27	0.03	0.22

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

Figure 1: **Pessimism and Financial Article Lengths**

This figure displays yearly arithmetic mean of article lengths and a corresponding sentiment measure constructed from daily data. Date (year in which the values are averaged) is on the x-axis, article length (*#Length*) or sentiment measure (*Pessimism*) are on the y-axes. Yearly arithmetic average of daily Dow log-returns, a known stationary variable, is included for comparison.



The inconsistent ADF and KPSS conclusions, along with the visually nonstationary behavior of the news-based variables, may also stem from the measure construction algorithm. It is possible that more advanced algorithms, such as those based on machine learning or artificial intelligence, could account for the non-stationarity of inputs. To investigate this possibility, I test multiple news-based measures that involve ML or AI at any stage of the construction process to determine whether these algorithms produce variables with statistical behavior similar to simple term frequencies. The results of this analysis are presented in Table 3.

I begin with a measure of daily news sentiment proposed in Shapiro et al. (2022). This measure uses an advanced, application-specific lexicon that assigns sentiment scores to domain-specific (economics/finance) terminology. Unlike simple term frequencies, the lexicon in Shapiro et al. (2022) incorporates "pointwise mutual information" accounting for the likelihood of an individual word influencing the overall sentiment of a sentence. The overall sentiment of a sentence is classified using VADER, an advanced machine learning algorithm specifically designed for sentiment analysis (see Hutto and Gilbert (2014)). Despite these methodological enhancements, this sentiment measure is stationary according to the ADF test but non-stationary according to the KPSS test, mirroring the results displayed by simple term frequencies.

Next, I evaluate a daily news-implied volatility measure (NVIX) proposed in Manela and Moreira (2017). NVIX is based on a support vector regression model, which uses daily news to predict the observed volatility (VXO index). The model is trained and tested on periods when both news and the VXO index are available, and it uses news alone to estimate volatility during periods when the VXO index is unavailable. NVIX is further decomposed into seven time series: government, financial intermediation, natural disasters, stock markets, war, and unclassified. Most of the news-implied volatility measures are stationary according to the ADF test but non-stationary according to the KPSS test. Specifically, six out of the seven time series, including the headline NVIX, show inconsistent results between the ADF and KPSS tests. For the natural disasters time series, the KPSS null hypothesis is also

rejected when the confidence level is relaxed from 1% to 2.5%.

Finally, I test the news-based measures of economic conditions spanning 180 topics proposed in Bybee et al. (2024). To automatically extract and model the topics, the authors use Latent Dirichlet Allocation, an advanced unsupervised machine learning technique. All daily measures (180/180) are stationary according to the ADF test. However, the KPSS test indicates that 178 out of 180 daily measures are nonstationary.[3] These unit root test conclusions are identical to those of the term counts, frequencies, and other news-based measures examined thus far.

Table 3: **Stationarity of Measures Constructed from Economic Texts**

This table shows the results of testing measures derived from financial texts using advanced machine learning or artificial intelligence methods for stationarity. A measure is labeled as stationary according to the ADF test in "Is Stationary?" column if all of the following null hypotheses are rejected at 1% level: presence of a unit root, unit root without trend, unit root without trend and without drift. A measure is labeled as non-stationary according to the KPSS test in "Is Stationary?" column if trend stationarity is rejected at 1% level. The number of lags for the ADF test is selected using AIC, and for the KPSS test it is set to $4(T/100)^{0.25}$. The results are robust to the lag selection.

	Is Stationary?		
Freq.	**ADF**	**KPSS**	**Measure Construction**
	News Sentiment from Shapiro et al. (2022), Jan. 1980 – Dec. 2024		
Daily	Yes, 1/1	No, 1/1	The measure is constructed using an application-specific dictionary.
	News-implied Volatility from Manela and Moreira (2017), Jul. 1889 – Mar. 2016		
Daily	Yes, 7/7	No, 6/7	The measure is constructed from term frequencies using a support vector regression to predict the actual observed volatility (VXO index).
	Business News, Multiple Topics from Bybee et al. (2024), Jan. 1984 – Jun. 2017		
Daily	Yes, 180/180	No, 178/180	Latent Dirichlet Allocation (LDA) is used to identify topics and construct measures. Topic identification is unsupervised and depends on word co-occurence. An individual measure is a fraction of an article dedicated to each topic. Monthly frequency is achieved by aggregating daily texts.
Monthly	Yes, 162/180	No, 160/180	

[3]Two variables, "Credit Cards" and "Changes", require relaxing the confidence level to 2.5% and 10%, respectively, for the KPSS test to reject the null hypothesis.

These measures of economic conditions are available at both daily and monthly frequencies, sharing the same relevant characteristics such as the construction algorithm, underlying source material, and time period. As a result, it is possible to isolate the effect of frequency on stationarity testing while keeping all other conditions unchanged. The ADF and KPSS tests yield inconsistent results for 142 out of 180 monthly measures. Specifically, 160 measures are identified as nonstationary by the KPSS test, while 162 are classified as stationary by the ADF test. Notably, 20 monthly measures are stationary according to both the ADF and KPSS tests – an outcome not observed for the daily news-based measures. The inconsistent unit root test outcomes likely arise from the combination of high frequency and text-based source material. After all, the ADF and KPSS tests are consistent at the daily frequency for conventional numeric measures such as Dow log-returns and some monthly text-based variables. On the other hand, measure construction algorithms and their level of sophistication (term frequency or ML/AI) do not seem to affect the stationarity of resulting variables.

Overall, it is likely that the combination of high frequency and the type of underlying source material (text) is responsible for the inconsistent ADF and KPSS test outcomes. It is also important to note that stationarity is rarely tested explicitly, even when assumed by the econometric model. In the rare instances when it is tested, researchers often rely solely on the ADF test. For example, Kalamara et al. (2022) present only the ADF test results. Ideally, all text-based measures should include visual evidence of stationarity and undergo testing with both the ADF and KPSS procedures. In the absence of such testing, economic conclusions – or any form of inference – should rely on statistics that are robust to the effects of nonstationarity.

2.3 Cointegration of Term Counts

Lexicons that classify terms by tone, such as those proposed in Loughran and McDonald (2011) and Garcia et al. (2023), are notably broad. For instance, consider the words "accident," "serious," "achieve," and "attain," along with their conjugations such as "accidental."

According to the Loughran and McDonald (2011) dictionary, "accident" and "serious" are categorized as negative, while "achieve" and "attain" are classified as positive. These terms – and many others like them – are very common and appear across a wide range of contexts, making them difficult to avoid unless deliberately excluded. Consequently, as the length of the text increases, the count of tonal terms inevitably rises. This observation suggests a shared underlying stochastic trend, i.e., a cointegrating relationship between term counts and text lengths. This cointegrating relationship is formally verified using Johansen (1991) and Phillips and Ouliaris (1990) tests; the results are presented in Table 4.

For all possible term count pairings (e.g., *#Negative* and *#Length*) and all specifications of the Johansen (1991) test, critical values far exceed the 1% threshold (7,694.29>16.26, 17,342.49>30.45, ..., 9,269.33>23.65), indicating the presence of at least one cointegrating vector. This conclusion is confirmed using the Phillips and Ouliaris (1990) procedure; similarly, the observed test statistics greatly exceed the 1% critical values (150,771.70>55.19, ..., 128,938.60>102.02). Statistically, these results confirm the existence of a long-term relationship between tonal term counts and text lengths. This finding is expected, especially considering the editorial process. The same author (or group of authors) writes newspaper articles subject to the publisher's constraints — primarily the word count of a column — which simultaneously limit both the count of tonal terms and the article lengths.

The cointegrating relationship between term counts and text lengths imposes limitations on the construction of text-based variables or their use in subsequent (second stage) economic models. For instance, term frequencies are often used as standalone variables, but simple division does not explicitly incorporate error correction or account for the cointegration between term counts and text lengths. Additionally, this cointegrating relationship can undermine the effectiveness (or even violate the assumptions) of some common econometric models used in conjunction with text-based variables. For example, regularized regression techniques such as LASSO, Ridge, and Elastic Net tend to perform poorly in the presence of multicollinearity. There are three potential solutions: adjust the construction procedure such that text-based variables satisfy the assumptions of existing methods, jointly estimate

the variables and the subsequent model, or employ robust statistics (or, when possible, bias correction) in the second stage. The rest of this paper focuses on robust statistics, as it is the only approach that does not require the reestimation of existing news-based variables.

Table 4: **Cointegrating Relationship Between Term Counts and Text Lengths**

This table reports the results of testing the individual term counts and text lengths (n=27,449 days) for cointegration using Johansen (1991) and Phillips and Ouliaris (1990) procedures. Both tests have a null hypothesis of no cointegration; alternative is a presence of cointegrating relationship. In Johansen (1991) test, "r" denotes a number of cointegrating vectors, critical values are from Osterwald-Lenum (1992), both trace and and eigenvalue-based tests are reported. P_z version of the Phillips and Ouliaris (1990) test is used, the input series are either unadjusted or demeaned with respect to a linear model with an intercept.

			Critical Values		
Variables	**Type**	**Test Statistic**	**10%**	**5%**	**1%**
		Johansen Test			
#Positive, #Length	Trace, $r \leq 1$	7,694.29	10.49	12.25	16.26
#Positive, #Length	Trace, $r = 0$	17,342.49	22.76	25.32	30.45
#Positive, #Length	Eigen, $r \leq 1$	6,546.38	10.49	12.25	16.26
#Positive, #Length	Eigen, $r = 0$	16,275.68	22.76	25.32	30.45
#Negative, #Length	Trace, $r \leq 1$	6,844.69	10.49	12.25	16.26
#Negative, #Length	Trace, $r = 0$	16,114.02	22.76	25.32	30.45
#Negative, #Length	Eigen, $r \leq 1$	7,694.29	10.49	12.25	16.26
#Negative, #Length	Eigen, $r = 0$	9,648.19	16.85	18.96	23.65
#Positive, #Negative	Trace, $r \leq 1$	6,546.38	10.49	12.25	16.26
#Positive, #Negative	Trace, $r = 0$	9,729.30	16.85	18.96	23.65
#Positive, #Negative	Eigen, $r \leq 1$	6,844.69	10.49	12.25	16.26
#Positive, #Negative	Eigen, $r = 0$	9,269.33	16.85	18.96	23.65
		Phillips–Ouliaris Test			
#Positive, #Length	P_z, Unadjusted	150,771.70	33.93	40.82	55.19
#Negative, #Length	P_z, Unadjusted	173,385.80	33.93	40.82	55.19
#Positive, #Negative	P_z, Unadjusted	156,394.40	33.93	40.82	55.19
#Positive, #Length	P_z, Demeaned	129,157.70	71.96	81.38	102.02
#Negative, #Length	P_z, Demeaned	142,314.70	71.96	81.38	102.02
#Positive, #Negative	P_z, Demeaned	128,938.60	71.96	81.38	102.02

3 Business Cycle and Return Predictability

3.1 Robust Inference

Creating variables from text, be it using simple term frequency or advanced AI/ML algorithms, is usually only the first stage of the analysis. More often, the focus is on uncovering a potential underlying relationship with other financial or economic variables. This relationship is usually investigated by employing a separate econometric model (second stage) where text-based variables are just one of the inputs and are treated exactly in the same way as conventional numeric data. However, due to their unknown statistical properties, including text-based variables may violate the assumptions behind the second-stage econometric model or invalidate statistical significance calculations.

Both of these issues are illustrated using the main result of Garcia (2013): "The predictability of stock returns using news' content is concentrated in recessions." This conclusion is drawn from the relative magnitudes of regression coefficients in recessions and expansions; statistical significance is based on the White (1980) standard errors. Nonetheless, the news-based sentiment measures are nonstationary at a daily frequency. In the worst case, nonstationarity may lead to nonsensical estimates of the regression coefficients or R^2 values, thereby challenging the economic conclusion. Furthermore, additional measurement error (e.g., sampling, text processing) from the first stage propagates into the second stage, leading to the imputed regressor problem (Pagan (1984), Murphy and Topel (1985)). Imputed regressors make OLS standard errors and common covariance adjustments (White (1980), Newey and West (1987), etc.) inapplicable; standard errors fail to account for the additional estimation uncertainty introduced in the first stage. From a financial perspective, this issue affects the interpretation of regression coefficients by making it impossible to determine whether a shock to the sentiment measure leads to a corresponding nonzero equity market response.

In this section, using the same nonstationary measures, I demonstrate that robust inference methods lead to the opposite findings. In fact, daily news sentiment forecasts stock

market returns at least as effectively in both recessions and expansions, with some evidence suggesting better predictability during expansions. The "All Controls" panel of Table 5 replicates Garcia (2013) and includes the estimates of the following model:

$$R_t = \beta_M M_{t-1} + Controls + \epsilon.$$

The controls ("*Controls*" in the formula above) include four additional lags of the sentiment measure ($M_{t-1}, ..., M_{t-5}$), five lags of Dow log-returns ($R_{t-1}, ..., R_{t-5}$), five lags of squared Dow log-returns ($R^2_{t-1}, ..., R^2_{t-5}$), and day-of-the-week indicators. The sentiment measures are normalized to have zero mean and unit variance. As a result, prediction betas can be interpreted as the market response to a one-standard-deviation sentiment shock.

Table 5: **Business Cycle and Return Predictability**

This table includes the estimates of $R_t = \beta_M M_{t-1} + Controls + \epsilon$ regressions. Full set of controls ("All Controls") is the same as in Garcia (2013) and include additional 4 lags of the sentiment measure ($M_{t-1}...M_{t-5}$), 5 lags of Dow log-returns ($R_{t-1}...R_{t-5}$), 5 lags of squared Dow log-returns ($R^2_{t-1}...R^2_{t-5}$), and day of the week indicators. Reduced set of controls includes one lag of Dow log-returns and squared Dow log-returns (R_{t-1}, R^2_{t-1}) while keeping all else the same. Statistical significance of β_M is computed using White (1980) procedure. MAE is mean absolute error; statistical difference between the MAE in recessions and expansions ("Rec. vs. Exp. MAE p-val.") is established using two-sample Welch's t-test. All p-values less than 0.001 are entered as 0.

	Recessions				**Expansions**				**Rec vs Exp**
	β_M **(bps)**	β_M ***p-val.***	***Adj.*R^2** **(%)**	**MAE (bps)**	β_M **(bps)**	β_M ***p-val.***	***Adj.*R^2** **(%)**	**MAE (bps)**	**MAE *p-val.***
	All Controls								
Positivity	7.75	0.0004	2.91	91.20	2.61	0.0002	1.78	64.10	0
Negativity	-6.71	0.01	2.86	91.01	-3.30	0	1.80	64.10	0
Pessimism	-9.73	0.0002	3.01	91.03	-4.07	0	1.85	64.09	0
	Reduced Controls								
Positivity	7.76	0.0005	0.74	91.44	2.35	0.001	0.88	64.27	0
Negativity	-6.70	0.01	0.68	91.28	-2.72	0.0002	0.90	64.26	0
Pessimism	-9.35	0.001	0.83	91.24	-3.41	0	0.94	64.26	0

The relative magnitudes of equity market prediction betas form a statistical basis for the economic conclusion in Garcia (2013): "The link between media content and Dow Jones Industrial Average (DJIA) returns is indeed concentrated in times of hardship." Specifically, the magnitude of β_M is larger in recessions than in expansions for all sentiment measures (Table 5, "All Controls": |7.75|>|2.61|, |–6.71|>|–3.30|, |–9.73|>|–4.07|). However, mean absolute errors (MAE) lead to the opposite economic conclusion. For all sentiment measures, MAE is consistently higher in recessions than in expansions, indicating worse predictive performance. Regardless of the variable, the average daily prediction errors are 91 bps in recessions and 64 bps in expansions; the difference is economically large and statistically significant (Table 5, "All Controls"). Crucially, unlike the regression beta estimates, MAE is a robust statistic that is unaffected by the nonstationarity of news-based sentiment measures.

Spurious regressions are sensitive to the choice of control variables and lag lengths. In such cases, even minor changes to the specification can lead to unstable estimates or inconsistent R^2 coefficients, reducing the reliability of economic analysis. Therefore, I introduce an alternative specification in which the number of lagged Dow log-returns and squared Dow log-returns is reduced to one while keeping all other aspects of the model unchanged (Table 5, "Reduced Controls"). Compared to the original, the R^2 coefficients in the alternative specification exhibit behavior inconsistent with the initial economic interpretation. In the original specification, adjusted R^2 coefficients are universally *higher* in recessions than in expansions for all sentiment measures (Table 5, "All Controls": 2.91>1.78, 2.86>1.80, 3.01>1.85). However, in the alternative specification, this relationship reverses; adjusted R^2 coefficients are universally *lower* in recessions (Table 5, "Reduced Controls": 0.74<0.88, 0.68<0.90, 0.83<0.94). Meanwhile, the robust statistic (MAE) behaves exactly as expected. MAE is consistently higher in recessions than in expansions (91>64) for all specifications and sentiment measures.

By employing robust inference and conservative significance thresholds, it is possible to add an extra layer of protection against the effects of a spurious regression without respecifying the model. To resolve the imputed regressor problem and further demonstrate the

influence of nonstationary news-based measures on parameter estimates, I use the approach described in Ibragimov and Muller (2010). This approach introduces a robust statistic, t_{IM}, which relies on the small-sample properties of Student's t-distribution to account for potentially heterogeneous data with an unknown correlation structure. The main idea behind t_{IM} can be informally stated as follows. For a significance level $\alpha \leq 0.083$, t-distribution with a sufficiently low number of degrees of freedom is so heavy-tailed that it absorbs additional uncertainty from the two-stage procedure. More formally, t_{IM} requires the data to be divided into a small number of groups (q), each representative of the full sample. The model of interest is then estimated separately for each group, resulting in $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1; \hat{\beta}_2; ...; \hat{\beta}_j\}$, where $j = 1, ..., q$.

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

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

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

H_{Null}: $\beta = \beta_0$ is then rejected in favor of H_{Alt}: $\beta \neq \beta_0$ if $|t_{IM}| > F_T^{-1}((1-\alpha/2), q-1)$, where $F_T^{-1}(p, df)$ is an inverse cumulative density function of the Student's t-distribution with df degrees of freedom. The statistic (t_{IM}) is asymptotically valid when the individual groups (q) are representative of the entire sample and the respective estimates ($\hat{\beta}_j$) are asymptotically independent (Ibragimov and Muller (2010)).

I group individual observations based on whether the year is divisible by q. For $q = 2$, the first group consists of all daily observations from even years ($year \equiv 0$ (mod 2)), while the second group includes all observations from odd years ($year \equiv 1$ (mod 2)). For example, the stock market return and the associated sentiment measurement observed on December 19, 1992 fall into group 1, whereas the same variables observed on December 19, 1993 belong to group 2. Model estimates using this arrangement are independent, yet representative of the entire sample. Both groups span the same year range, and the individual

daily observations are included in continuous, uninterrupted year-long blocks that retain the original autocorrelation structure. The groups are defined similarly for $q = 4$: $year \equiv 0$ $(mod\ 4)$, $year \equiv 1$ $(mod\ 4)$, $year \equiv 2$ $(mod\ 4)$, $year \equiv 3$ $(mod\ 4)$.

According to Ferson et al. (2003), spurious relationships and their effects are more pronounced in small samples. By construction, t_{IM} requires the sample to be divided into smaller groups, thus increasing the likelihood of observing a disproportionate $\hat{\beta}_j$ estimate. A low number of degrees of freedom $(q-1)$ does not allow any of $\hat{\beta}_j$ to deviate much from $\hat{\beta}_{Avg}$ while retaining statistical significance. In other words, an outlier among $\hat{\beta}_j$ would have a large effect on t_{IM}. For example, consider the estimate of the response of the equity market to a change in pessimism in expansions ($\beta_{Pessimism} = -4.07$) and its statistical significance (t_{IM}=–3.02, $q = 4$). The estimate is not statistically significant at either the 1% or 5% level; $t_{IM} = -3.02$ falls inside both the 1% (± 5.84) and 5% (± 3.18) confidence intervals. The associated $\hat{\boldsymbol{\beta}}$ is $\{-2.985099;\ -8.423142;\ -2.302426;\ -3.295377\}$. In this case, the second group estimate, $\hat{\beta}_{j=2} = -8.423142$, is considerably different from the rest. With only $q - 1 = 3$ degrees of freedom, the outlier has a substantial effect on inference, rendering the t_{IM} statistic insignificant.

Table 6 provides t_{IM} for all estimates in Table 5. Specifically, the $\beta_M = 0$ null hypothesis is tested against a $\beta_M \neq 0$ alternative. In recessions, none of the β_M estimates are significant at the 1% level, regardless of whether $q = 2$ or $q = 4$. When $q = 2$, $t_{IM} = 30.99, -2.51, -4.14, 14.24, -2.72, -3.63$ (depending on the sentiment measure), all of which fall within the ± 63.66 confidence interval. Similarly, when $q = 4$, the associated $t_{IM} = 3.74, -3.14, -3.72, 3.29, -3.06, -3.24$, all of which fall within the ± 5.84 confidence interval. This lack of statistical significance marks a stark departure from the White (1980) p-values in Table 5, which are universally below 0.01. Taken together, these statistically insignificant t_{IM} values provide evidence against the economic conclusion in Garcia (2013). It is difficult to argue that news-based sentiment predicts equity market returns in recessions better than in expansions when the corresponding beta estimates are not statistically significantly different from zero.

Table 6: **Robust Inference: Magnitude of Dow Response to Change in Sentiment**

This table includes the results of testing $\beta_M = 0$ against $\beta_M \neq 0$ hypothesis using robust Ibragimov and Muller (2010) t-statistic (t_{IM}). β_M comes from $R_t = \beta_M M_{t-1} + Controls + \epsilon$ regression. M is a sentiment measure. Full set of controls ("All Controls") is the same as in Garcia (2013) and include additional 4 lags of the sentiment measure ($M_{t-1}...M_{t-5}$), 5 lags of Dow log-returns ($R_{t-1}...R_{t-5}$), 5 lags of squared Dow log-returns ($R^2_{t-1}...R^2_{t-5}$), and day of the week indicators. Reduced set of controls includes one lag of Dow log-returns and squared Dow log-returns (R_{t-1}, R^2_{t-1}) while keeping all else the same. $cval, 1\%$ and $cval, 5\%$ are critical values associated with the $\alpha/2 = .005$ and $\alpha/2 = .025$ confidence intervals respectively.

	t_{IM} **(2 Groups)**				t_{IM} **(4 Groups)**			
	Rec	**Exp**	$cval, 1\%$	$cval, 5\%$	**Rec**	**Exp**	$cval, 1\%$	$cval, 5\%$
	All Controls							
Positivity	30.99	9.66	±63.66	±12.71	3.74	11.01	±5.84	±3.18
Negativity	-2.51	-2.07	±63.66	±12.71	-3.14	-2.24	±5.84	±3.18
Pessimism	-4.14	-2.85	±63.66	±12.71	-3.72	-3.02	±5.84	±3.18
	Reduced Controls							
Positivity	14.24	7.94	±63.66	±12.71	3.29	13.06	±5.84	±3.18
Negativity	-2.72	-1.68	±63.66	±12.71	-3.06	-2.11	±5.84	±3.18
Pessimism	-3.63	-2.32	±63.66	±12.71	-3.24	-2.92	±5.84	±3.18

Although not directly shown here, when chosen appropriately, robust statistics are also valuable during the variable creation in the first stage. For example, t_{IM} can be paired with AI/ML techniques that rely on partitioning the sample into smaller subsamples to assess whether the model fit is sufficiently uniform during cross-validation (a machine learning technique where a model is trained using subsets of the data and tested against an independent sample not used in the training process). The statistic is computationally efficient, so incorporating it into the first stage would not require additional resources. The main drawback of t_{IM} is the requirement to partition the input data into smaller groups, each representative of the full sample. Implicitly, this is a sample size requirement: the full sample needs to be sufficiently large to allow for such grouping. However, AI/ML methods are typically deployed when the volume of data is so large that it cannot be processed efficiently by other means, making the drawback inconsequential from the practical perspective.

3.2 Asymmetric Equity Market Volatility

Second-stage model specification often depends on the economic and statistical properties of the inputs. At the same time, the properties of text-based variables created in the first stage are generally unknown. As a result, interactions between variables in the second stage can be misattributed; this is even more likely when employing a black-box methodology. I directly show that equity market volatility, which is much higher during recessions than during expansions, affects both forecast errors and regression beta estimates. Specifically, the volatility and the spurious correlation between equity returns and sentiment jointly explain the difference in regression beta magnitudes. Most importantly, higher prediction errors correspond to periods of elevated volatility. An additional econometric model is required to account for the time variation in volatility and isolate the connection between news sentiment and Dow returns. After employing a GARCH(1,1) model, equity return forecast errors are statistically equal during recessions and expansions across all sentiment measures.

Table 7 presents two sets of predictive regressions: a full model with all controls ("All Controls") and a model without any controls ("Only Sentiment"). The "Only Sentiment" model includes only one explanatory variable, a news-based sentiment measure ($R_t = \beta_M M_{t-1} + \epsilon$). The "Only Sentiment" model also retains all statistically and economically significant properties of the "All Controls" model. The magnitudes of prediction betas are consistently larger during recessions than during expansions for all sentiment measures (Table 7, "Only Sentiment": |8.80|>|3.43|, |–7.67|>|–3.94|, |–9.96|>|–4.67|). Additionally, MAE is indistinguishable between the "Only Sentiment" and "All Controls" specifications. Regardless of the specification, the average prediction error is approximately 91 bps during recessions and 64 bps during expansions (Table 7). In effect, the magnitudes of prediction betas and forecast errors are unaffected by the choice of controls; they depend only on the economic state and the choice of sentiment measure.

This statistical similarity implies that the conclusions drawn from the analysis of a simpler "Only Sentiment" model would very likely translate to the full specification. The "Only Sentiment" specification is analytically simple, and prediction betas can be easily decomposed

Table 7: **Model Specification and Return Predictability Statistics**

This table shows that the estimates of equity market return predictability are unaffected by the model specification and the choice of controls. The table includes the estimates of $R_t = \beta_M M_{t-1} + Controls + \epsilon$ ("All Controls") and $R_t = \beta_M M_{t-1} + Controls + \epsilon$ ("Only Sentiment") regressions. Full set of controls ("All Controls") is the same as in Garcia (2013) and include additional 4 lags of the sentiment measure ($M_{t-1}...M_{t-5}$), 5 lags of Dow log-returns ($R_{t-1}...R_{t-5}$), 5 lags of squared Dow log-returns ($R^2_{t-1}...R^2_{t-5}$), and day of the week indicators.

	All Controls				**Only Sentiment**			
	Recessions		**Expansions**		**Recessions**		**Expansions**	
	$\boldsymbol{\beta_M}$	**MAE**	$\boldsymbol{\beta_M}$	**MAE**	$\boldsymbol{\beta_M}$	**MAE**	$\boldsymbol{\beta_M}$	**MAE**
Positivity	7.75	91.20	2.61	64.10	8.80	91.64	3.43	64.45
Negativity	-6.71	91.01	-3.30	64.10	-7.67	91.40	-3.94	64.43
Pessimism	-9.73	91.03	-4.07	64.09	-9.96	91.37	-4.67	64.42

into underlying components without any cross-terms. Such decomposition helps identify the conditional (on the state of the business cycle) contribution of each component to the relative difference in the magnitude of prediction betas. According to the definition of regression beta for a $R_t = \beta_M M_{t-1} + \epsilon$ model (R_t are Dow log-returns and M is a news-based sentiment measure),

$$\hat{\beta}_M = \frac{Cov(Measure, Dow)}{Var(Measure)} = \hat{\rho}_{M,Dow} \frac{\hat{\sigma}_{Dow}}{\hat{\sigma}_{Measure}}.$$

Three factors determine the magnitude of the prediction beta: the correlation between the news-based sentiment measure and the Dow log-returns ($\hat{\rho}_{M,Dow}$), the volatility of the equity market, and the volatility of a sentiment measure. Importantly, the realized volatility of Dow log-returns is completely independent of the news-based sentiment and is notably higher during recessions.

For example, consider the headline news-based sentiment measure, *Pessimism*, and the

corresponding prediction betas: –9.96 bps during recessions and –4.67 bps during expansions.

$$\hat{\beta}_{Pessimism,Dow} = -0.06645\frac{0.01410}{0.9473} = -9.958\ bps\ (recessions)$$

$$\hat{\beta}_{Pessimism,Dow} = -0.05052\frac{0.00938}{1.0146} = -4.671\ bps\ (expansions)$$

This decomposition shows that the estimate of equity market volatility ($\hat{\sigma}_{Dow}$) accounts for most of the difference in prediction betas between recessions and expansions. The difference in correlation coefficients is influenced by the spurious relationship between news-based sentiment and Dow log-returns. The volatility of the sentiment measure itself does not deviate significantly over the business cycle (0.9473 during recessions, 1.0146 during expansions). Moreover, the volatility of the sentiment measure is primarily determined by the change in the count of tonal terms from one daily newspaper to the next. The term count is constrained by the column length, which is largely independent of economic conditions. On the other hand, the standard deviation of the Dow log-returns during recessions is 1.41%, 1.5 times higher than in expansions (0.938%). As a result, the asymmetry in prediction betas ($|\beta_{M,Recession}| > |\beta_{M,Expansion}|$) is also observed for the *Positivity* (8.80>3.43, Table 7) and *Negativity* (|–7.67|>|–3.94|, Table 7). Stated alternatively, the magnitude of the prediction beta is determined by equity market return volatility, independently of news-based sentiment or the choice of sentiment measure.

Importantly, the correlation coefficient between the sentiment measure and equity market returns ($\hat{\rho}_{M,Dow}$) is solely responsible for the sign of the prediction beta and the corresponding economic interpretation of the relationship. However, the correlation coefficients between the news-based sentiment measures and Dow log-returns are not economically significant and do not vary with the business cycle (Table 8). For example, the correlation between *Positivity* and Dow log-returns is 0.043 for the entire sample, 0.058 during recessions, and 0.037 during expansions. For *Negativity*, the respective values are –0.045, –0.051, and –0.043. For *Pessimism*, the corresponding values are –0.055, –0.066, and –0.051. The difference in

magnitudes between recessions and expansions is insignificant and is affected by the spurious relationship. The correlation coefficients are also very low and close to zero.

It is unlikely that the influence of daily sentiment on equity market returns is persistent. Newspaper articles written long ago (e.g., 10 years ago) have no impact on the market today, yet the historical relationship between sentiment and returns influences the estimate of the correlation coefficient. Figure 2 displays 5-year (5*252=1,260 daily observations) rolling correlation between sentiment measures (*Positivity*, *Negativity*, *Pessimism*) and daily Dow log-returns. Visually, for all measures, the rolling correlation coefficient neither fluctuates around a fixed mean nor has a discernible trend. The sign of these correlation coefficients also flips from positive to negative depending on the time period. Taken altogether, this structural instability indicates that the estimate of a long-term relationship between news-based sentiment and equity market returns is economically meaningless.

Table 8: **Correlations between Sentiment Measures and Dow Returns**

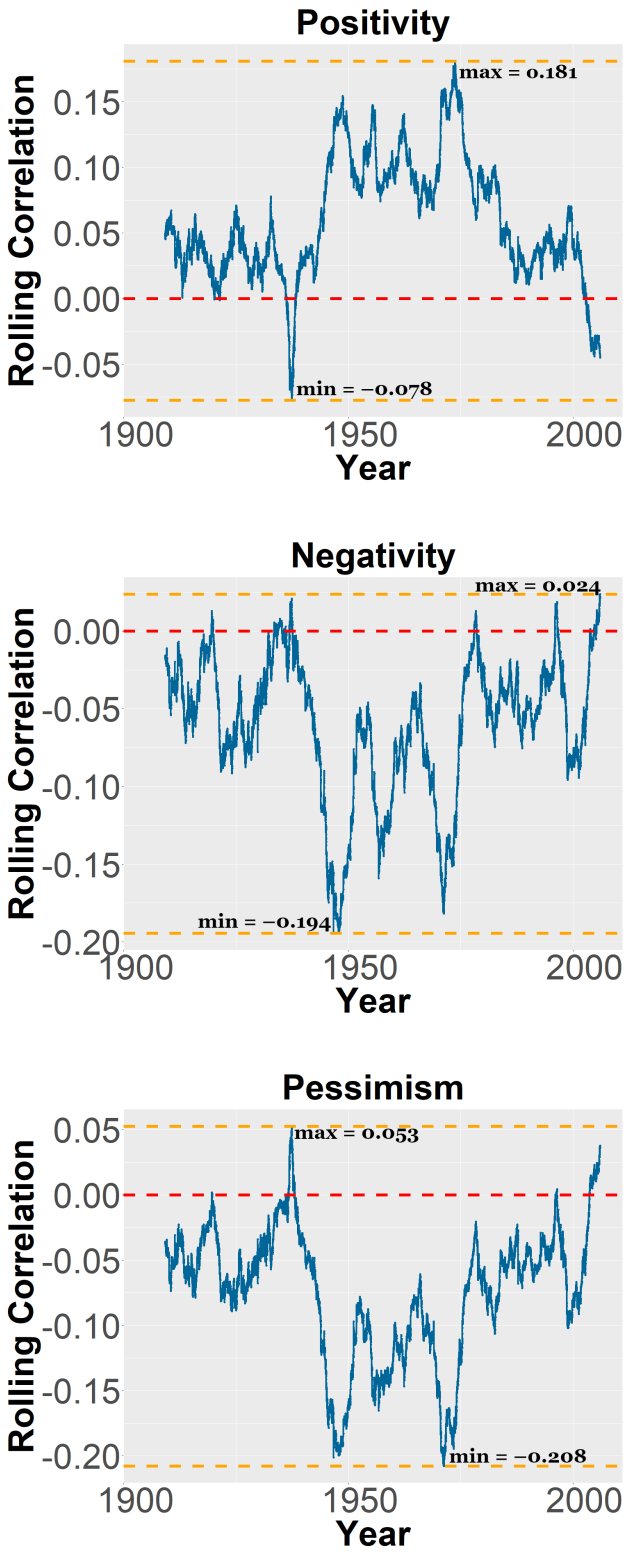
This table includes pairwise correlations between the news-based sentiment measures (*Positivity*, *Negativity*, and *Pessimism*) and daily Dow log-returns (Dow), n=27,449. Correlations during recessions (n=6,455) and expansions (n=20,994) are shown on the bottom panel above and below the diagonal respectively. Recession indicator is NBER USRECD.

	All Dates			
	Pos	**Neg**	**Pess**	**Dow**
Positivity	1	-0.253	-0.673	0.043
Negativity	-0.253	1	0.886	-0.045
Pessimism	-0.673	0.886	1	-0.055
Dow	0.043	-0.045	-0.055	1
	Recessions and Expansions			
	Pos	**Neg**	**Pess**	**Dow**
Positivity	1	-0.264	-0.673	0.058
Negativity	-0.250	1	0.891	-0.051
Pessimism	-0.673	0.884	1	-0.066
Dow	0.037	-0.043	-0.051	1
	Expansions			

Recessions

Figure 2: **5-Year Rolling Correlation between Sentiment Measures and Dow**

This figure displays 5-year (5*252=1,260 daily observations) rolling correlation between sentiment measures (*Positivity*, *Negativity*, *Pessimism*) and daily Dow log-returns.



3.2.1 Heteroskedastic Residuals and MAE

The difference in volatility between expansions and recessions results in unequal, state-dependent variance of the residuals. In an OLS setting, this state-dependent variance affects only the standard errors of the estimates, not the parameters themselves. However, this difference in volatility directly influences robust residual-based statistics. Without accounting for it, mean absolute errors during recessions (a more volatile state) are generally higher than during expansions. More formally, conditional heteroskedasticity of the residuals introduces bias into MAE estimation, leading to unreliable economic inference.

To account for the difference in volatility between expansions and recessions, I follow Garcia (2013) and use a GARCH(1,1) process to estimate the conditional variance of daily returns. GARCH(1,1) performs well in a variety of financial settings (Hansen and Lunde (2005)) and is a deterministic, heavy-tailed process (Bollerslev (1986)). As a result, it accounts for volatility clustering and is also immune to data snooping. Then, to construct a time series of normalized equity market returns, I rescale R_t by the GARCH(1,1) estimate of conditional standard deviation ($\hat{\sigma}_t$). By construction, the resulting volatility-adjusted return time series ($Adj.R_t$) has unit variance.

$$\begin{aligned} Ret_t &= \mu_t + \epsilon_t; \;\; \sigma^2_{t+1} = \omega + \alpha_1 \epsilon^2_t + \beta_1 \sigma^2_t \\ \sigma^2_t &\equiv Var(\epsilon_t) \\ Adj.R_t &= R_t / \hat{\sigma}_t \end{aligned}$$

I proceed by estimating the same predictive regression but replace the dependent variable, Dow log-returns, with the volatility-adjusted time series ($Adj.R_t = \beta_M M_{t-1} + \epsilon$). The results are presented in Table 9. Judging from the MAE estimates, the predictive performance of news-based sentiment is the same during recessions and expansions, regardless of the measure. For example, the MAE estimates obtained using the headline *Pessimism* measure are 74.82 and 74.63 bps in recessions and expansions, respectively. The difference between these estimates is not statistically significant (p-value = 0.84).

These estimates differ from the results obtained with the unadjusted Dow log-returns, where the heteroskedasticity of residuals directly influences the MAE. The resulting economic interpretation is also different: news-based sentiment predicts the volatility-adjusted equity market return time series at least equally well (or poorly) during expansions and recessions. This interpretation is subjective: it depends on whether asymmetric volatility is a market property inseparable from the returns or a separate characteristic subject to an econometric correction. In either case, the results here are contrary to Garcia (2013). Daily news-based sentiment forecasts stock market returns at least as effectively during both recessions and expansions, with some evidence suggesting better predictability during expansions. This example highlights the importance of knowing the properties of the inputs (both economic and statistical) and then choosing an appropriate methodology. If the methodology includes black-box AI/ML algorithms, verifying the results, preferably externally, becomes essential to ensure the validity of economic inference.

Table 9: **Volatility-Adjusted Mean Absolute Error in Recessions and Expansions**

This table presents MAE from predicting volatility-adjusted daily Dow Jones log-returns using $Adj.R_t = \beta_M M_{t-1} + \epsilon$ regression. MAE during recessions is compared to MAE during expansions; p-values are calculated using two-sample Welch's t-test; p-values less than 0.0001 are entered as zero. Volatility-adjusted daily returns are obtained by fitting GARCH (1,1) model with a constant mean ($Ret_t = \mu_t + \epsilon_t$) and time-varying volatility ($\sigma^2 = \omega + \alpha_1 \epsilon_t^2 + \beta_1 \sigma_t^2$, $\sigma_t^2 \equiv Var(\epsilon_t)$). Adjusted returns are $Adj.R_t = R_t / \hat{\sigma}_t$, where $\hat{\sigma}_t$ is estimated in the previous step. GARCH(1,1) parameter standard errors and t-statistics are robust.

	MAE (bps)		
	Rec	**Exp**	***p-value***
Positivity	75.03	74.67	0.70
Negativity	74.84	74.64	0.83
Pessimism	74.82	74.63	0.84
	GARCH(1,1) Parameters		
	Est.	**SE**	***t-stat***
$\boldsymbol{\mu}$	0.0004	0.0001	5.356
$\boldsymbol{\omega}$	0	0	0.040
$\boldsymbol{\alpha_1}$	0.066	0.046	1.430
$\boldsymbol{\beta_1}$	0.939	0.041	23.141

4 Informational Content of News Sentiment

4.1 In-Sample

It remains uncertain whether news-based sentiment measures contain additional information beyond what is already captured by prior equity market returns. To address this, I conduct a series of tests to assess the informational content (or lack thereof) of these measures and to demonstrate that the seemingly asymmetric predictability arises directly from state-dependent equity market volatility. The analysis builds on the empirical framework employed in Welch and Goyal (2008) and Goyal et al. (2024), and focuses on comparing the predictive performance of sentiment measures to that of simple benchmarks derived solely from historical returns.

Table 10 presents the results of the in-sample testing. Consistent with the previous sections of this paper, news-based sentiment measures (*Positivity*, *Negativity*, and *Pessimism*) are used to forecast daily Dow Jones log-returns using the $R_t = \beta_M M_{t-1} + \epsilon$ model. Two simple benchmarks, derived solely from the equity market returns, are used for the evaluation: historical average and one-day momentum. The historical average benchmark (labeled "Cnst") is a $R_t = \beta_{Const.} 1 + \epsilon$ regression, and the one-day momentum ("Lag") is a $R_t = \beta_{R_{t-1}} R_{t-1} + \epsilon$ model. Benchmark-relative predictive performance is determined by comparing the mean absolute errors (MAE) of the sentiment measures to those of the benchmarks. Once again, I use both unadjusted and volatility-adjusted daily returns (scaled using the GARCH (1,1) model) to demonstrate the influence of the volatility asymmetry between expansions and recessions. The results are reported in the "MvsC" column (sentiment measure vs. the historical average) and the "MvsL" column (sentiment measure vs. the one-day momentum benchmark).

It is immediately clear that the predictive performance of the news-based measures is statistically indistinguishable from that of the benchmarks (Table 10). All sentiment measures exhibit predictive performance comparable to both benchmarks, regardless of the state of the business cycle or the return time series (with or without volatility adjustments). For

example, when predicting equity market returns with the *Pessimism* measure, the MAE during recessions is 91.37 bps for the unadjusted returns, and 74.82 bps after incorporating the volatility adjustments. Under the same conditions, the historical average benchmark MAEs are 91.67 bps and 75.07 bps, while the one-day momentum benchmark MAEs are 91.63 bps and 75.03 bps. The corresponding p-values, 0.87, 0.89, 0.83, and 0.86, show that there is no statistically or economically significant difference in predictive performance between the sentiment measures and the benchmarks. Importantly, the benchmark strategies exhibit the same asymmetric MAE behavior as the sentiment measures, with the asymmetry disappearing after the volatility adjustments. This suggests that the observed asymmetric predictive performance is a feature of the market itself, driven by the differences in volatility between expansions and recessions. Alternatively stated, the asymmetric MAE behavior is not a characteristic of the sentiment measures, but rather a result of market dynamics.

Table 10: **Benchmark-Relative In-Sample Return Predictability**

This table compares the in-sample predictability of daily Dow Jones log-returns (with and without volatility adjustments) between news-based sentiment measures and benchmarks derived from market returns. News-based sentiment measures are *Positivity*, *Negativity*, and *Pessimism*; the equity market return forecasting relies on a $R_t = \beta_M M_{t-1} + \epsilon$ regression. The first benchmark is a historical average, $R_t = \beta_{Const.} 1 + \epsilon$ (labeled "Cnst"). The second benchmark is one-day momentum, $R_t = \beta_{R_{t-1}} R_{t-1} + \epsilon$ ("Lag"). Benchmark-relative predictive performance is determined by comparing the sentiment mean absolute errors (MAE) to the benchmark MAE. The results are reported in "MvsC" (sentiment measure against the historical average benchmark) and "MvsL" (sentiment measure against the one-day momentum benchmark) columns; p-values are calculated using two-sample Welch's t-test. Volatility-adjusted daily returns are obtained by fitting GARCH (1,1) model with a constant mean ($Ret_t = \mu_t + \epsilon_t$) and time-varying volatility ($\sigma^2 = \omega + \alpha_1 \epsilon_t^2 + \beta_1 \sigma_t^2$, $\sigma_t^2 \equiv Var(\epsilon_t)$). Adjusted returns are $Adj.R_t = R_t / \hat{\sigma}_t$, where $\hat{\sigma}_t$ is estimated in the previous step.

	Recessions					**Expansions**				
	MAE (bps)			***p-value***		**MAE (bps)**			***p-value***	
	Meas	**Cnst**	**Lag**	**MvsC**	**MvsL**	**Meas**	**Cnst**	**Lag**	**MvsC**	**MvsL**
	Unadjusted Dow Returns									
Positivity	91.64	91.67	91.63	0.99	0.99	64.45	64.47	64.41	0.98	0.95
Negativity	91.40	91.67	91.63	0.89	0.91	64.43	64.47	64.41	0.95	0.97
Pessimism	91.37	91.67	91.63	0.87	0.89	64.42	64.47	64.41	0.94	0.99
	Volatility-Adjusted Dow Returns									
Positivity	75.03	75.07	75.03	0.97	1.00	74.67	74.69	74.56	0.98	0.86
Negativity	74.84	75.07	75.03	0.84	0.87	74.64	74.69	74.56	0.93	0.90
Pessimism	74.82	75.07	75.03	0.83	0.86	74.63	74.69	74.56	0.92	0.91

4.2 Out-of-Sample

Look-ahead bias is a potential issue in in-sample evaluation and, in this context, it may significantly overstate the predictive performance of benchmarks. While forward-state information is incorporated into the benchmark beta estimates ($\beta_{Const.}$ and $\beta_{R_{t-1}}$), sentiment measures are constructed using only prior data. Additionally, time series length influences in-sample evaluation. Since beta estimates incorporate both historical and recent data, this may understate the predictive performance of news-based sentiment. For example, during the Great Depression, word-of-mouth was the primary source of information, as only a select few could afford newspapers. Yet, the estimate of the relationship is carried over throughout the evaluation sample. Economically, this issue is inapplicable only if the sentiment is persistent, a claim empirically challenged in this paper (see Figure 2 and the corresponding discussion).

To address the issues with in-sample evaluation, I also assess the benchmark-relative out-of-sample predictive performance. I set the evaluation window to 1,260 trading days, roughly corresponding to five calendar years (5*252=1,260), and then incrementally re-estimate the predictive regressions and the benchmark strategies (historical average and one-day momentum). Following Welch and Goyal (2008) and Goyal et al. (2024), I evaluate out-of-sample predictive performance using $R^2_{OOS} = 1 - MSE_{Meas.}/MSE_{Bench.}$, where MSE represents mean squared error. Results are presented in Table 11.

Judging from Table 11, it is readily apparent that benchmark-relative predictive performance does not differ between recessions and expansions. For example, without volatility adjustments, *Positivity*, *Negativity*, and *Pessimism* slightly outperform the historical average benchmark during recessions, posting R^2_{OOS} values of 0.38%, 0.40%, and 0.54%, respectively. However, the corresponding R^2_{OOS} values during expansions are 0.25%, 0.32%, and 0.45%. The differences between predictive performance during recessions and expansions is neither economically nor statistically significant. Interestingly, none of the news-based sentiment measures outperform the one-day momentum benchmark strategy; all corresponding R^2_{OOS} values are negative and, once again, statistically and economically indistinguishable between

recessions and expansions. This empirical finding further confirms that sentiment is not persistent; even a short 5-year window is not competitive with a one-day benchmark strategy. Incorporating volatility adjustments does not affect the results; there is still no difference in predictive performance between recessionary and expansionary periods.

Table 11: **Out-of-Sample Return Predictability in Recessions and Expansions**

This table compares the out-of-sample Dow Jones log-return (with and without volatility adjustments) predictability between news-based sentiment measures and benchmarks derived from market returns. News-based sentiment measures are *Positivity*, *Negativity*, and *Pessimism*; the equity market return forecasting relies on an incrementally re-estimated $R_t = \beta_M M_{t-1} + \epsilon$ regression over a fixed 5-Yr.*252=1,260 daily observations window. The first benchmark is a historical average calculated by incrementally re-estimating $R_t = \beta_{Const.} 1 + \epsilon$ regression (labeled "Constant"). The second benchmark is a momentum strategy calculated by incrementally re-estimating $R_t = \beta_{R_{t-1}} R_{t-1} + \epsilon$ regression ("Lag"). Benchmark-relative predictability is indicated by $R^2_{OOS} = 1 - MSE_{Meas.}/MSE_{Bench.}$, where MSE is mean squared error. Volatility-adjusted daily returns are obtained by fitting GARCH (1,1) model with a constant mean ($Ret_t = \mu_t + \epsilon_t$) and time-varying volatility ($\sigma^2 = \omega + \alpha_1 \epsilon_t^2 + \beta_1 \sigma_t^2$, $\sigma_t^2 \equiv Var(\epsilon_t)$). Adjusted returns are $Adj.R_t = R_t/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is estimated in the previous step.

	Out-of-Sample R^2 (%)			
	Constant		**Lag**	
	Recessions	**Expansions**	**Recessions**	**Expansions**
	Unadjusted Dow Returns			
Positivity	0.38	0.25	-0.61	-0.72
Negativity	0.40	0.32	-0.60	-0.65
Pessimism	0.54	0.45	-0.46	-0.52
	Volatility-Adjusted Dow Returns			
Positivity	0.55	0.56	-0.86	-0.94
Negativity	0.99	0.58	-0.42	-0.91
Pessimism	1.21	0.85	-0.19	-0.65

Overall, there is very limited evidence that news-based sentiment measures exhibit asymmetric predictive performance depending on the state of the business cycle. Moreover, there is little to no empirical evidence that news-based sentiment measures constructed using term frequencies are useful for predicting daily market returns; they are unlikely to outperform the benchmark strategies derived from historical market returns.

4.3 Validation

In this section, I validate previous findings by demonstrating that they are not specific to the time period, are not an artifact of the variable construction algorithm, and are replicable using alternative news-based measures of sentiment. Specifically, I employ the daily news sentiment index published by the San Francisco Fed (Shapiro et al. (2022)) and show that it predicts Dow Jones log-returns (with and without volatility adjustments) *worse* during recessions than during expansions. This index represents a stark departure from basic term frequency calculations. Instead, it relies on an advanced, application-specific lexicon that scores domain-specific (economics/finance) terminology and accounts for the likelihood of an individual word influencing the overall sentiment of a sentence. As part of its construction process, sentence-level sentiment is assessed using VADER, a specialized rules-based sentiment analysis tool.

The San Francisco Fed daily news sentiment index covers the period from 1980 to the present. The news-based sentiment measures employed throughout this paper cover 1905 to 2005. As a result, I consider two separate (though overlapping) time periods: 1980–2024 (the complete SF Index time series) and 1980–2005 (the period common to the SF Index, *Positivity*, *Negativity*, and *Pessimism*). Predictive performance (MAE) is then evaluated in-sample since, as demonstrated above, the results do not differ from those computed out-of-sample. Consistent with the rest of this paper, I use both unadjusted daily Dow Jones log-returns and log-returns scaled by the GARCH(1,1) model. To further assess the results and provide an additional layer of robustness, I also include the previously employed benchmark strategies: historical average and one-day momentum. Table 12 shows the results.

First, the complete SF Index time series (1980–2024) replicates the results previously obtained in this paper using the *Positivity*, *Negativity*, and *Pessimism* sentiment measures. The SF Index predicts daily Dow Jones log-returns less accurately during recessions than during expansions, both with and without volatility adjustments. The MAE is 118.05 bps during recessions and 68.11 bps during expansions without volatility adjustments, and 81.25 bps and 73.65 bps, respectively, with volatility adjustments. The difference is statistically and

economically significant. Similarly, the predictive performance of the benchmark strategies mirrors that of the sentiment index and is worse during recessions. Both the historical average and one-day momentum benchmarks have MAEs that are indistinguishable from those of the SF Index. The errors are approximately 118 bps during recessions and 68 bps during expansions without volatility adjustments, and 81 bps and 73 bps, respectively, with adjustments.

Table 12: **San Francisco Fed News Sentiment Index and Return Predictability**

This table compares the predictability of daily Dow Jones log-returns (with and without volatility adjustments) during recessions and expansions using the San Francisco Fed news sentiment index (Shapiro et al. (2022), labeled "SF Index"). Predictive performance (MAE) is computed for two time periods: 1980-2024 (complete SF Index time series; n=11,296, 1,207 recessionary days) and 1980-2005 (overlap with the term frequency measures; n=6,551, 788 recessionary days). MAE during recessions is compared to MAE during expansions; p-values are calculated using two-sample Welch's t-test; p-values less than 0.0001 are entered as zero. The results are presented alongside the term frequency sentiment measures (*Positivity*, *Negativity*, *Pessimism*) and two benchmarks derived from market returns. The first benchmark is a historical average, $R_t = \beta_{Const.} 1 + \epsilon$ (labeled "Constant"). The second benchmark is one-day momentum, $R_t = \beta_{R_{t-1}} R_{t-1} + \epsilon$ ("Lag"). Volatility-adjusted daily returns are obtained by fitting GARCH (1,1) model with a constant mean ($Ret_t = \mu_t + \epsilon_t$) and time-varying volatility ($\sigma^2 = \omega + \alpha_1 \epsilon_t^2 + \beta_1 \sigma_t^2$, $\sigma_t^2 \equiv Var(\epsilon_t)$). Adjusted returns are $Adj.R_t = R_t / \hat{\sigma}_t$, where $\hat{\sigma}_t$ is estimated in the previous step. GARCH (1,1) model parameters are estimated separately for each time period.

	Unadjusted Dow MAE (bps)			**Volatility-Adjusted Dow MAE (bps)**		
	Rec	**Exp**	***p-value***	**Rec**	**Exp**	***p-value***
			1980-2024			
SF Index	118.05	68.11	0	81.25	73.65	0.0004
Constant	118.02	68.11	0	81.25	73.65	0.0004
Lag	117.39	68.12	0	81.35	73.68	0.0003
			1980-2005			
SF Index	85.18	71.28	0	78.09	74.08	0.13
Constant	85.16	71.27	0	78.08	74.08	0.13
Lag	85.16	71.28	0	78.12	74.21	0.13
Positivity	85.20	71.31	0	78.13	74.13	0.13
Negativity	85.05	71.35	0	77.95	74.15	0.15
Pessimism	85.10	71.37	0	78.01	74.18	0.14

The second time period under consideration (1980–2005) allows me to directly compare all relevant predictors: the SF Index, historical average and one-day momentum benchmark strategies, and all news-based sentiment measures (*Positivity*, *Negativity*, and *Pessimism*). Nevertheless, the results remain the same. Regardless of the predictor, MAE without volatility adjustments is approximately 85 bps during recessions and 71 bps during expansions. With volatility adjustments, MAE during recessions and expansions is approximately 78 bps and 74 bps, respectively. For all combinations of predictors, time periods, and return time series, forecasting performance is worse during recessions – an empirical finding that directly contradicts the conclusion of Garcia (2013).

Daily news sentiment forecasts stock market returns at least as effectively during both recessions and expansions, with some evidence suggesting better predictability during expansions. These empirical findings are generally consistent with two different strands of economic models: noise trading (Shleifer and Summers (1990)) and investor disagreement (Hong and Stein (2007)). Noise trading models rely on the absence of disagreement among investors. Noise traders influence asset prices when their demand shifts are correlated; that is, when a noisy signal is interpreted in the same way by many traders. Asymmetric predictability would suggest that noisy signals are more correlated during one phase of the business cycle. However, arbitrageurs, who act as counterparties to noise traders, are active during both recessions and expansions, neutralizing any potential asymmetric demand shifts. As a result, news-based signals unrelated to market fundamentals (i.e., sentiment) are equally ineffective in both recessions and expansions and, therefore, possess no forecasting power. Furthermore, the noise trading model in Yu and Yuan (2011) suggests that noise traders are more likely to actively participate in the market when the outlook is positive, which helps explain why, in some periods, news-based sentiment performs better during expansions.

Investor disagreement models provide an even simpler explanation for the symmetric equity market return predictability – or, more precisely, the lack thereof – during both recessions and expansions. For example, a negative headline such as "Investors Haven't Been This Pessimistic About Stocks Since 2023" (Wall Street Journal, Feb. 17, 2025) may

trigger opposite trades. For some, it signals an opportunity to buy the dip, while for others, it prompts to cash out. Positive news similarly lead to heterogeneous opinions and actions. For some market participants (e.g., momentum traders), a positive headline is a signal to buy, whereas for others (e.g., mean reversion traders), it creates pressure to lock in profits. As long as the degree of disagreement remains independent of the business cycle, news will have neither predictive nor explanatory power.

5 Conclusion

Machine learning methods receive considerable criticism for being opaque. This issue runs deeper: both the data fed into these algorithms and the resulting output often have unknown statistical properties. In this sense, unstructured data, such as text, is as much a "black box" as the advanced algorithms used to process it. Additionally, variables constructed from unstructured data are rarely the ultimate goal. Instead, these measures are used as inputs in subsequent economic models, which also typically incorporate conventional numeric data. Interactions between heterogeneous variables remain largely unexplored in economic settings, potentially leading to incorrect inference. This paper has two main contributions: documenting the undesirable statistical properties of news-based variables that may adversely impact economic inference, and demonstrating how existing robust statistical methods can be used in conjunction with these measures to produce reliable conclusions.

I illustrate incorrect economic inference arising from interactions between text-based and conventional variables by revisiting Garcia (2013) study, which focuses on predicting equity market returns during recessions and expansions using daily news sentiment measures. Using robust statistical methods, I demonstrate that the main empirical finding of Garcia (2013), "the predictability of stock returns using news' content is concentrated in recessions," is based on a spurious relationship and an unaccounted-for interaction with the economic properties of equity market returns. In fact, daily news sentiment forecasts stock market returns at least as effectively during both recessions and expansions, with some evidence suggesting

better predictability during expansions.

I document that high-frequency news-based variables and their underlying building blocks, such as text lengths and term counts, are often nonstationary, regardless of the construction methodology. Nonstationarity is also difficult to detect; the high-frequency behavior of the ADF and KPSS tests is often inconsistent, with the KPSS test generally being more reliable than the ADF. Moreover, individual term counts and text lengths are often cointegrated – a property that variable construction algorithms (especially AI/ML) or subsequent economic models may fail to take into account.

I propose two complementary approaches to address these issues. First, it is crucial to pre-test text-based variables to ensure that they satisfy the assumptions of the subsequent econometric models. Second, economic inference should be based on statistics robust to data heterogeneity and unknown correlation structure. Such statistics provide an additional layer of safety and are more likely to result in Type II error than Type I. In the context of equity return predictability, residual-based robust statistics (such as mean absolute error and robust t-statistic introduced in Ibragimov and Muller (2010)), along with benchmarking, are sufficient to ensure reliable economic inference. However, other financial and economic applications may require different statistical procedures, depending on the models and research questions.

It is also important to note that unstructured data can be of higher quality than conventional data. For example, Martinez (2022) finds that "autocracies overstate yearly GDP growth by approximately 35%" by examining the deviations between self-reported GDP and night-time luminosity. Nevertheless, it is not always possible to compare results relying on unstructured data with those obtained solely using conventional numeric measures. The data might not be measured frequently enough (with inflation expectations and sentiment being classic examples) or may be withheld for a variety of reasons. Notably, Russia and China have a track record of stopping the publication of unfavorable or politically sensitive economic data, as illustrated by the recent invasion of Ukraine ("Russia Blocks Economic Data, Hiding Effect of Western Sanctions," The Wall Street Journal, Apr. 23, 2022) and

a spike in youth unemployment ("China Slashes Rates, Suspends Youth Jobless Data as Economy Signals Sharper Downturn," The Wall Street Journal, Aug. 15, 2023). There may simply be no alternatives to alternative data; a statistical safety layer that addresses the undesirable properties of unstructured data is a step toward reliable economic inference.

References

Elliott Ash and Stephen Hansen. Text Algorithms in Economics. *Annual Review of Economics*, 15(1):659–688, 2023.

Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for Regression with Variables Generated by AI or Machine Learning. *Working Paper*, 2024. URL https://arxiv.org/abs/2402.15585.

David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 04 1986. https://doi.org/10.1016/0304-4076(86)90063-1.

Leland Bybee, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. Business News and Business Cycles. *The Journal of Finance*, 79(5):3105–3147, 2024. https://doi.org/10.1111/jofi.13377.

David A. Dickey and Wayne A. Fuller. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74:427–431, 06 1979. 10.2307/2286348.

Wayne E. Ferson, Sergei Sarkissian, and Timothy T. Simin. Spurious Regressions in Financial Economics? *The Journal of Finance*, 58:1393–1413, 07 2003. 10.1111/1540-6261.00571.

Diego Garcia. Sentiment during Recessions. *The Journal of Finance*, 68:1267–1300, 2013.

Diego Garcia, Xiaowen Hu, and Maximilian Rohrer. The colour of finance words. *Journal of Financial Economics*, 147(3):525–549, 2023. ISSN 0304-405X. https://doi.org/10.1016/j.jfineco.2022.11.006.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as Data. *Journal of Economic Literature*, 57:535–574, 09 2019. 10.1257/jel.20181020.

Amit Goyal, Ivo Welch, and Athanasse Zafirov. A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 37(11):3490–3557, 2024.

Peter R. Hansen and Asger Lunde. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20:873–889, 2005. https://doi.org/10.1002/jae.800.

Harrison Hong and Jeremy C Stein. Disagreement and the Stock Market. *Journal of Economic Perspectives*, 21:109–128, 04 2007. 10.1257/jep.21.2.109.

C. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–225, May 2014. 10.1609/icwsm.v8i1.14550.

Rustam Ibragimov and Ulrich K. Muller. t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics*, 28:453–468, 10 2010. 10 .1198/jbes.2009.08046.

Bibo Jiang, Ye Lu, and Joon Y. Park. Testing for Stationarity at High Frequency. *Journal of Econometrics*, 215(2):341–374, 2020. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom .2019.09.004.

Soren Johansen. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59:1551, 11 1991. 10.2307/2938278.

Eleni Kalamara, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia. Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37, 06 2022. 10.1002/jae.2907.

Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 10 1992. 10.1016/0304-4076(92)90104-y.

Tim Loughran and Bill McDonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66:35–65, 01 2011. 10.1111/j.1540-6261 .2010.01625.x.

Asaf Manela and Alan Moreira. News implied volatility and disaster concerns. *Journal of Financial Economics*, 123:137–162, 01 2017. 10.1016/j.jfineco.2016.01.032.

Luis R Martinez. How Much Should We Trust the Dictator's GDP Growth Estimates? *Journal of Political Economy*, 130(10):2731–2769, 2022.

Kevin M. Murphy and Robert H. Topel. Estimation and Inference in Two-Step Econometric Models. *Journal of Business & Economic Statistics*, 3:370, 10 1985. 10.2307/1391724.

Whitney K. Newey and Kenneth D. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55:703, 05 1987. 10.2307/1913610.

Michael Osterwald-Lenum. 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, 08 1992. 10.1111/j.1468-0084.1992.tb00013.x.

Adrian Pagan. Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, 25:221–247, 1984.

P. C. B. Phillips and S. Ouliaris. Asymptotic Properties of Residual Based Tests for Cointegration. *Econometrica*, 58:165, 1990. 10.2307/2938339.

Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243, 2022. ISSN 0304-4076. https://doi.org/10.1016/j.jeconom.2020.07.053.

Andrei Shleifer and Lawrence H Summers. The Noise Trader Approach to Finance. *Journal of Economic Perspectives*, 4(2):19–33, 1990.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

Ivo Welch and Amit Goyal. A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21:1455–1508, 2008.

Halbert White. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48:817, 05 1980. 10.2307/1912934.

Jianfeng Yu and Yu Yuan. Investor sentiment and the mean–variance relation. *Journal of Financial Economics*, 100(2):367–381, 2011.

Guofu Zhou. Measuring Investor Sentiment. *Annual Review of Financial Economics*, 10: 239–259, 11 2018. 10.1146/annurev-financial-110217-022725.