Large Language Models and Financial Market Sentiment [*]

Shaun A. Bond, Hayden Klok, and Min Zhu

UQ Business School
The University of Queensland

June 15, 2024

Abstract

We investigate the predictive capabilities of large language models (LLMs) in the context of forecasting aggregate stock market returns. We use ChatGPT to analyse daily news summaries of the U.S. stock market, from which we construct a market sentiment indicator for the S&P 500 Index. Our findings reveal a noteworthy negative correlation between this sentiment indicator and short-term market returns. Notably, LLMs outperform conventional sentiment classifiers, with ChatGPT exhibiting a slight edge in out-of-sample performance. This analysis underscores the substantial potential of LLMs in text analysis — a relatively underexplored data source — for gaining insights into asset markets.

[*] We would like to thank the following for comments and feedback: Emanuele Bajo, Vincent Mellor, Sterling Yan, Qi Zeng, Rumi Masih, Eric Tan, and participants of the 2023 FIRN Asset Management Meeting. Contact authors: shaun.bond@uq.edu.au. h.klok@uq.edu.au min.zhu@uq.edu.au

1 Introduction

Recent years have seen a growing number of financial studies on the role of investor sentiment and asset pricing. One area that continues to draw particular attention is the extraction of information from text. Motivated by behavioural models of noise traders and mispricing (De Long et al., 1990), prior literature has shown valuable financial insights can be gained through text analysis (Tetlock, 2007; Garcia, 2013). Research has demonstrated better quantification of investor sentiment can provide mispricing insights, illuminate the cross section of stock returns, and predict short-term return reversals. These studies serve to highlight the information richness of text and its relevance to asset pricing.

While the integration of text-based information in financial modelling is a growing area, working with text comes with unique challenges, as it is often unstructured and contains mixed signals and topics. Given these obstacles, traditional text analysis methods used simple dictionary approaches, however these lack context-awareness, which is an important aspect of text as it can help clarify financial communication (Garcia et al., 2023; Tetlock et al., 2008). More recently, advances in machine learning techniques that can better account for semantic meaning and context have extended the literature on investor sentiment and asset pricing, with recent studies suggesting transformer neural networks and language models can better capture contextual information and sentiment than previous methods, leading to enhanced prediction in the cross section of stock returns (Jiang et al., 2022; Ke et al., 2019).

Our paper extends this literature, as we use large language models (LLMs) such as ChatGPT and BARD to derive proxies for aggregate U.S. stock market sentiment from historical end of day financial news text summaries, and then use these sentiment measures in a series of models that forecast future returns of the S&P 500 Index. We compare the performance of model forecasts using LLM-derived sentiments against models that incorporate sentiments derived from traditional dictionary and simpler transformer text analysis methods, and find that ChatGPT outperforms traditional dictionary approaches. We explore the potential underlying economic channels behind our results by investigating sentiment significance over different periods of higher and lower; economic uncertainty, economic policy uncertainty, times of recession, and business conditions. Consistent with prior literature, results show sentiment reversals are strongest and more significant when uncertainty is greatest and market conditions are worse, suggesting the effects of news sentiment on investor behaviour is strongest during these times, (Tetlock, 2007). In addition, we investigate the ability of LLMs ChatGPT and BARD to recall historical financial information over time, and demonstrate similar performance to models that incorporate sentiments proxied from news.

We make three contributions to the literature. First, we demonstrate ChatGPT is capable of classifying sentiment from news text from the perspective of a financial advisor. Second, we show these sentiments forecast future aggregate market return reversals in line with behavioural models. Third, we demonstrate ChatGPT and BARD can provide text summaries of aggregate market-level financial news. Our results show the incorporation of sentiment improves model performance, and incorporating LLM derived sentiments leads to greater outperformance compared to models using sentiments from simpler transformers or traditional dictionary approaches.

The recent advent of LLMs has been particularly noteworthy, with their impressive ability to perform a variety of text-based tasks.[1] Currently, the two most well-known and widely used LLMs are OpenAI's Generative Pre-trained Transformer ChatGPT, and Google Gemini - previously called BARD.[2] While there are mixed views and concerns on the use of LLMs, (Zaremba and Demir, 2023; Zuckerman, 2023), recent financial studies have shown LLMs can be used to predict stock returns (Lopez-Lira and Tang, 2023), classify 'Fedspeak' (Hansen and Kazinnik, 2023), identify information in conference call Q&A's (Bai et al., 2023), analyse sentiment of Japanese stock news (Nakano and Yamaoka, 2023), provide financial advice (Fieberg et al., 2023), help with portfolio stock selection (Romanko et al., 2023), and even have a level of financial literacy (Niszczota and Abbas, 2023). These studies serve to underscore the potential use of LLMs in financial research and practice.

One natural application of LLMs is in helping to illuminate aspects of behavioural finance through the analysis of financial texts.[3] There has been a growing number of studies on investor sentiment in recent years, including proxying investor sentiments from fundamentals (Baker and Wurgler, 2006), web searches (Da et al., 2015), and news images (Obaid and Pukthuanthong, 2022), to measuring investor attention (Da et al., 2011), pessimistic news (Tetlock, 2007), sentiment during uncertain times (Birru and Young, 2022), and sentiment and uncertainty (Baker et al., 2016). It

[1] Fundamentally LLMs are huge transformer neural networks which estimate a general language representation from a massive text corpora. Given some text input, they respond with an answer based on this learned language generalisation. Hence LLMs differ from traditional transformers trained to perform singular tasks, and one conceivable advantage of LLMs is they have the potential to better identify textual semantics and context than simpler transformer models or algorithms.

[2] OpenAI. (2023). *OpenAI API (gpt-3.5-turbo)*. https://platform.openai.com/docs/api-reference. Accessed: 1 August 2023; Google. (2024). *Gemini*. https://blog.google/products/gemini/bard-gemini-advanced-app. Accessed: 15 March 2024; Google. (2023). *BARD*. https://bard.google.com/. Accessed: 1 August 2023.

[3] Behavioural finance models posit two key assumptions. First, some investors are irrational and prone to a variety of biases, with some of these investors able to affe

is in this context that LLMs represent an exciting new tool that may aid in the identification of textual information and sentiment, given their potential for enhanced text processing and contextual understanding.

The recent works of Lopez-Lira and Tang (2023), Jiang et al. (2022), and Chen et al. (2023) are most closely related to our study. Lopez-Lira and Tang (2023) use ChatGPT to classify U.S. stock news headlines from RavenPack as either long, short, or uncertain, and then use these sentiment scores to predict daily stock returns. Jiang et al. (2022) derive measures of sentiment for individual stocks from Reuters news articles using several simpler transformer neural networks, the open-source language model OPT, (Zhang et al., 2022), and several dictionary methods. They find that portfolios based on transformer language model sentiments outperform portfolios that incorporate sentiments from simpler transformers and dictionary methods. Although they do not use the LLMs ChatGPT or BARD, they suggest ChatGPT has possible future use as a sentiment analyser. Finally, the parallel work of Chen et al. (2023) use ChatGPT to identify if markets will go up or down based on news headlines and alerts from the Wall Street Journal, and after aggregating responses, investigate the return predictability of the U.S. stock market.

Our study differs from these papers in several aspects. First, we focus on overall market index return predictability, and aggregate market-level sentiment rather than the cross section of stock returns. In addition, we provide ChatGPT with longer text input for the purpose of sentiment classification via end-of-day market summary texts, and do not aggregate sentiment over multiple items. In supplementary work, we also uniquely explore the potential of ChatGPT and BARD to recall historical news text summaries, and classify sentiments from these recalled texts. Results demonstrate the inclusion of sentiments from LLM-recalled text leads to similar performance to that of news-based sentiments, and suggest ChatGPT outperforms BARD in forecasting returns out-of-sample. Interestingly, weekly and monthly results suggests daily text summaries from both LLMs recall information that is persistent at these lower frequencies, while news-based sentiments do not. In the context of sentiments derived from news, this behaviour supports the premise that news reflects contemporaneous events, and that news, and news sentiments, contribute little new information to market prices, as reversals are driven by initial investor over-reaction, which later corrects. In addition, these results suggest LLMs may contain lookahead bias when recalling historical events.

The rest of the paper is structured as follows. Section 2 discusses our methodology and sentiment measures. Section 3 outlines the modelling and evaluation procedure. Section 4 details our main results. Section 5 examines underlying economic channels. Section 6 discusses LLM text recall. Section 7 concludes.

2 Sentiment Construction

2.1 Daily News Summary Text

LLMs can perform a variety of tasks, including question answering, recalling historical information, and summarising text. Hence in the future LLMs could be used to assist financial analysts process text from sources such as financial news articles, analyst reports, and conference call transcripts. To derive proxies of daily investor sentiment we use web-scraped end-of-day U.S. stock market news summaries from Reuters over a 20-year period from 1-1-2000 to 31-7-2020. We rely on daily news text summaries of the U.S. stock market as our study focuses primarily on the daily returns of S&P 500 index, and we believe the sentiments derived are a good proxy for the overall investor sentiment at the time. Summary statistics are presented in Table 1 and Figure 1.

	Daily News Summary Text	
	Characters	Words
Total	9,582,924	1,361,444
Mean	1,864	265
Std dev	2,071	326
Minimum	86	15
Maximum	19,141	3,191
No. days	5,142	

Table 1: Summary statistics for Reuters end-of-day U.S. stock market summary text.

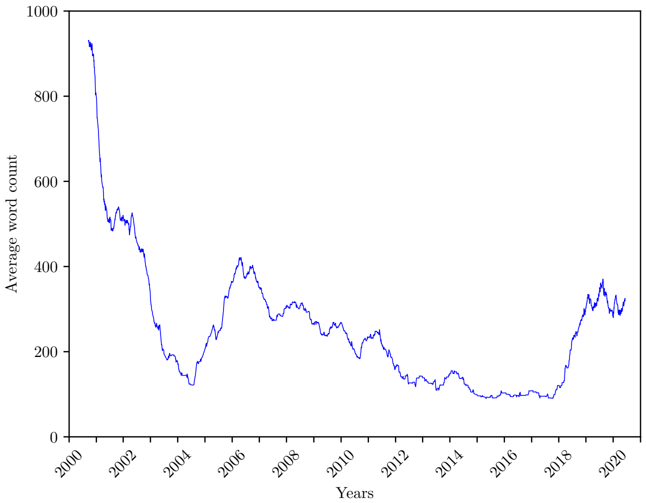


Figure 1: 180-day moving average word counts for U.S. stock market news summaries from 1-1-2000 to 31-7-2020.

While the length of market summary text varies over time, the average hovers around 200 words. Generally, news peaks around times of greater market volatility, coinciding with events such as the September 11 terrorist attacks, the Global Financial Crisis, and COVID-19. The large standard deviation in length is primarily driven by several days early in the sample with particularly long text content. The reason for these lengthier days is unclear, but may be due to different policies or methodology towards news reporting at the start of the sample. There were also a handful of days with particularly short, one line summaries. Nevertheless, in general the news summaries are much longer than short one sentence news headlines used in several other studies. The remainder of this section details how we derive sentiments from these daily texts.

2.2 ChatGPT Sentiment Construction

To interact with ChatGPT we use OpenAI's paid-for API service. The two most used models at the time were gpt-3.5-turbo-0301 (gpt-3.5) and DaVinci-003. We use gpt-3.5, based on OpenAI's recommendation that it outperforms DaVinci on tasks such as text summarisation and sentiment classification, while being ten times cheaper. Per OpenAI's documentation, we set the 'temperature' hyperparameter to zero, to ensure model responses were as deterministic and replicable as possible, noting however that some randomness remains since ChatGPT models are non-deterministic by nature. We leave all other settings as default.

To extract measures of ChatGPT sentiment (GP), we start a new API session, and submit summary text for each day as separate input, while prompting the LLM to classify the sentiment of the text. To help align ChatGPT's answers towards relevant responses, we append the following text to the start of each days summary text:

"Act as a financial advisor. On a scale of 1 to 100, where 100 is the highest and 1 is the lowest, separately evaluate a score of positivity, neutrality, and negativity from the prompt text below. Make sure to evaluate each class separately and independently. Do not round your answers to whole numbers, instead provide granular answers to one decimal point. Do not include any explanatory or clarifying text, output the ratings only. The text to evaluate is: TEXT"

The responses consist of scores of positivity, neutrality, and negativity, along with some additional text and formatting. We parse the text, and take the scores as our daily ChatGPT sentiment measures.[4]

2.3 Traditional Sentiment Construction

Alongside ChatGPT, we use four other text classifiers to derive measures of positive and negative sentiment tone from the daily news summaries. We use two traditional dictionary-based approaches - Loughran & McDonald and VADER, and two transformer neural networks trained for sentiment classification - FinBERT and TwitterRoBERTa.[5]

The traditional dictionary methods of Loughran & McDonald and VADER use a set of human-curated words or rules to classify text, and hence may be more robust in a low signal-to-noise environment (Jiang et al., 2022). However, one of their main drawbacks is they are unable to identify contextual information or derive semantic meaning from text. By comparison, deep neural network models such as ChatGPT, FinBERT and TwitterRoBERTA, can take into account semantic meaning and context through their neural network architectures. Rather than relying on a set of human-curated rules, these models encode words into numeric representations or tokens, and from these word embeddings are able to perform tasks such as text classification. Hence they represent a different approach to text classification than dictionary methods.

[4] We use the score % by dividing by 100. See Appendix B.1 for parsing details.

[5] FinBERT and TwitterRoBERTa were run locally on two Nvidia GTX 1080 graphics cards, reducing runtime by a factor of 20.

At their core, FinBERT and TwitterRoBERTA are based on similar transformer neural network architecture to LLMs such as ChatGPT and BARD. However they were developed several years prior, can only perform sentiment classification of text, and are much smaller, with FinBERT and TwitterRoBERTa having 110M and 125M parameters respectively, while gpt-3.5 has 175B parameters. We include both FinBERT and TwitterRoBERTA as they represent an advancement beyond simple dictionary classification methods, and allow us to compare the performance of ChatGPT to other classifiers that have a degree of context awareness. To differentiate FinBERT and TwitterRoBERTA from truly large language models, we refer to them as transformers.

By including traditional dictionary and transformer models alongside ChatGPT, we aim to better compare the performance of sentiment classification of text. As with ChatGPT, we separately classify each days summary text. The remainder of this section discusses each classifier in more detail.[6]

2.3.1 Loughran & McDonald Dictionary

The Master Dictionary of Loughran and McDonald (2011) (LM) is a list of positive and negative words specifically curated for use in financial context. We use this dictionary, along with a negation check of positive words to count the number of positive and negative words on each day.[7] Negation is based on a third list of 'negate' keywords, which change a positive word to a negative word if a negate keyword is observed within three preceding words of a positive word.[8] Following Loughran and McDonald's recommendation, this negation check was only applied to positive words, since double negation is not common. We take each days proportion of positive and negative words over total daily words as the daily positive and negative sentiments.

2.3.2 VADER

The VADER (VA) sentiment analyser by Hutto and Gilbert (2014) uses a human-curated lexicon and grammatical ruleset to take into account word and sentence intensity and context to classify sentiment from text. We use VADER from Python's Natural Language Toolkit, along with the default

[6] Although several similar BERT-based transformers for sentiment classification exist, we do not use BERT or other binary classifiers with only two classes positive and negative, as the resulting classification probabilities are collinear. FinBERT and TwitterRoBERTa have a third neutral class, and hence do not suffer as greatly from positive-negative collinearity.

[7] Chen, K. (2019). *Use Python to calculate the tone of financial articles.* Kai Chen. http://kaichen.work/?p=399. Accessed: 12 May 2023.

[8] See Appendix A.5.

‘vader_lexicon’. While the default lexicon can be updated to include domain-specific vocabulary or lexical features, we use the default unmodified lexicon, as it represents an early attempt to improve on simple dictionary word count approaches, and has been shown to be applicable across several different domains. We run VADER on each days text, which outputs several ‘polarity_scores’. The first three represent probabilities the text is classed as negative, neutral, or positive, with values between [0,1] which sum to 1. The fourth output is an overall sentiment measure of the text, a value between [-1,1] negative to positive. We take the daily positive and negative polarity scores as positive and negative VADER sentiments.

2.3.3 FinBERT

The transformer neural network FinBERT (FB) developed by Araci (2019) is available via the open source machine learning community Hugging Face (Wolf et al., 2020). A modified version of the BERT language model (Devlin et al., 2018), FinBERT was developed by further training BERT on a subset of financial text from Reuters, specifically TRC2-financial, and then fine-tuning the model on the human labelled Financial PhraseBank (Malo et al., 2014) for the purpose of classifying sentiment tone of financial text.[9]

Note around the same time, a different FinBERT model was developed in a separate work by Yang et al. (2020). Also based on the BERT transformer, this model was further trained on a financial text corpus consisting of corporate reports, earnings call transcripts, and analyst reports, for the purpose of sentiment classification. However, there are reported inconsistencies in model outputs from the web API and locally run transformer model instances, with seemingly erroneous answers observed, such as output probabilities greater than 1.[10] Due to these concerns, we use the (Araci, 2019) model.

One point of note is FinBERT input text must be less than 512 tokens, or roughly 300 words in length, due to internal dimensions of the transformer. Hence we make use of the ‘truncation=True’ option in the FinBERT pipeline to automatically truncate text if this limit is reached. This is a common approach when input text is close to the maximum length. Since our aim is overall text classification, and since most daily text summaries are within this limit, we expect only minor loss

[9] National Institute of Standards and Technology (U.S.), (2018). *Reuters Corpora TRC2.* https://doi.org/10.7910/DVN/IEJ2UX.

[10] See discussions on the Huggingface model page discussing inconsistent model results as recently as July 2023, which are yet to be addressed.

of information, if any.

FinBERT's default output consists of a single classification label; positive, neutral, or negative, along with a softmax score between [0,1]. Rather than use this single score, we take the corresponding logits from the previous model layer for two main reasons. First, the softmax (final) layer only returns the probability of the maximum class, hence we lose information from the other two. Second, the softmax probabilities are bound between [0,1], while the logits are bound $[-\infty, \infty]$ and hence allow for better relative comparison between days. Hence we use the positive and negative logits as our measures of FinBERT sentiment.

2.3.4 TwitterRoBERTa

TwitterRoBERTa (TR) (Barbieri et al., 2020), available via Hugging Face, is a RoBERTa-based BERT model which has been fine-tuned for sentiment analysis on approximately 58 million Twitter tweets with the TweetEval benchmark.[11] We include this transformer as it has been fine-tuned for sentiment analysis on a completely different type of text, namely social media. Because of this the same text may be evaluated in similar, but subtly different ways compared to FinBERT, and this offers us a second way to compare the sentiment classification of ChatGPT.

TwitterRoBERTa has the same input token limit as FinBERT, and manual truncation is performed since there is no option to truncate input text in the pipeline automatically.[12] The output is the same as FinBERT, and the positive and negative logits are taken as the daily measures of positive and negative sentiment.

2.4 Text Sentiment Summary

Summary statistics of sentiment measures for daily U.S. stock market news summaries is presented in Table 2. Overall, sentiments across all classifiers are similar. On average, dictionary methods report more negative sentiments than positive. The transformers FB and TR have the most granularity due to logits used, while GP is the least granular, as responses were sometimes rounded to the nearest 5 or 10% despite our prompt engineering instructions. Further, GP return values at both

[11] Twitter rebranded to 'X' in July 2023. https://twitter.com/elonmusk/status/1683171310388535296

[12] Input text is encoded, truncated to 512 tokens, decoded back to text, and the remaining text classified.

the lower and upper limits of it's range, while in contrast LM and VA have low maximum values due to the proportion of positive and negative words present. FB and TR have relatively higher maximum values than minimum for both positive and negative sentiment, with the exception of negative sentiment for TR.

	LM		VA		FB		TR		GP	
	+	-	+	-	+	-	+	-	+	-
Mean	0.012	0.023	0.072	0.046	-0.337	1.150	-0.141	-0.887	0.450	0.305
Median	0.010	0.018	0.070	0.037	-0.845	2.204	-0.454	-0.638	0.500	0.200
P25	0.000	0.009	0.049	0.017	-1.790	-1.006	-1.582	-2.363	0.100	0.040
P75	0.017	0.032	0.094	0.069	1.323	2.982	1.187	0.653	0.800	0.400
Std dev	0.013	0.019	0.038	0.038	1.489	1.935	1.540	1.646	0.319	0.303
Minimum	0.000	0.000	0.000	0.000	-1.984	-2.390	-2.458	-3.783	0.010	0.010
Maximum	0.088	0.117	0.225	0.222	2.052	3.109	3.199	1.998	1.000	0.999

Table 2: Summary statistics of daily positive and negative sentiment measures for daily U.S. stock market news summaries. The five sentiment classifiers are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), and ChatGPT (GP). Results are rounded to 3 decimal places, and are not winsorized or standardized.

Generally, negative sentiments are greater in magnitude than positive sentiments, suggesting that either more intense negative signals are present, or extreme negative sentiments are being identified more strongly, consistent with prior literature of bad news being stronger than good (Baumeister et al., 2001). Most sentiments have positive skew, and ChatGPT has higher variance than the other classifiers. This could indicate that ChatGPT is capturing a wider range of context-based information, and hence have a wider range of positive and negative sentiments.

Next we compare daily sentiments with contemporaneous daily % returns of the S&P 500 Index. Figure 2 shows Pearson correlation coefficients between sentiment measures and contemporaneous daily returns. Results show positive (negative) sentiments correlate to positive (negative) same day returns, consistent with existing literature of news-based sentiments reflecting references to contemporaneous events. GP has the strongest correlations to same-day returns, followed by FB and TR, with traditional dictionary methods LM and VA last.

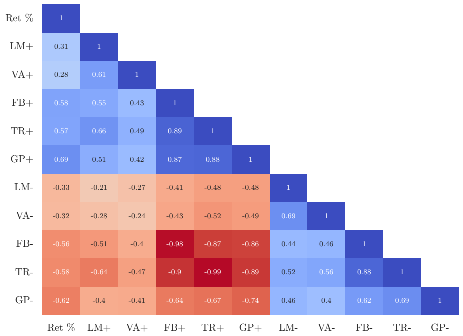


Figure 2: Sentiment correlations of daily U.S. stock news text summaries, along with daily S&P 500 returns.

Comparing sentiments with each other, there is strong and consistent positive correlations between sentiments of the same tone, and inverse correlations between opposing tones. FB and TR have the strongest correlations, likely as they hail from the same original BERT transformer model. GP has the next strongest correlations, and dictionary methods generally have the weakest. Overall, these results indicate ChatGPT can classify financial text sentiment in a similar manner to traditional methods.

3 Return Prediction Model and Evaluation

3.1 Prediction Models

To investigate how sentiments relate to future market returns we run a panel of OLS regression models that fit next day returns using one day lagged sentiments from each sentiment classifier separately. The model is:

$$R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}, \quad (1)$$

with,

$$Sent_{t,C} = C_{t(-)} - C_{t(+)} \quad \text{for} \quad C \in \{LM,\ VA,\ FB,\ TR,\ GP\}.$$

That is, $Sent_{t,C}$ is the overall daily sentiment for classifier C, calculated as the negative sentiment measure minus the positive sentiment measure for classifier C, one of $\{LM,\ VA,\ FB,\ TR,\ GP\}$, per Section 2. Some studies in the literature use both positive and negative sentiment in regressions, (Tetlock, 2007; Heston and Sinha, 2017), acknowledging potentially different effects from both positive and negative news signals. We use a single overall sentiment measure for two reasons. First, negative sentiment signals and associated reversals are generally larger than positive sentiment signals (Garcia, 2013). Second, by using a single sentiment measure in the regression model we remove potential collinearity issues that could occur by including two opposing sentiment variables. Although LM's positive and negative sentiments are derived independently, this is not the case for the other classifiers. Figure 2 shows high correlation between several positive and negative sentiment pairs, for example FB and TR. Hence by using a single measure, we reduce the chance of potential model instability.

Alongside sentiment, we control for several variables which potentially predict returns. The control variables are:

R_t: S&P 500 Index total % returns, excluding dividends, data from Refinitiv Datastream (Refinitiv, 2021b).

VIX_t: Chicago Board Options Exchange Volatility Index VIX divided by 100, data from Refinitiv Datastream/Eikon (Refinitiv, 2021b).

$Tedrate_t$: short-term credit risk. For dates prior to 21 January 2022 is the daily spread between 3-Month USD LIBOR rate [USD3MTD156N] and 3-Month Treasury Bill rate [DTB3]. For dates after 21 January 2022, is the daily spread between the daily Secured Overnight Financing Rate [SOFR] and 3-Month Treasury Bill rate [DTB3], due to discontinuation of the [TEDRATE] series. Data from (Federal Reserve Bank of St. Louis., 2023d), and (Federal Reserve Bank of New York., 2023).[13]

$Termspread_t$: medium-term credit risk, calculated from the daily spread between 10-Year Treasury Bond rates [T10YFF], and 1-Year Treasury Bond rates [T1YFF]. Data from (Federal Reserve Bank of St. Louis., 2023a).

$BAA10Y_t$: long-term credit risk, is the daily spread between Moody's Seasoned Baa Corporate Bond yield averages, based on bonds with maturities 20 years and above [DBAA], and 10-Year Treasury Constant Maturity [BC_10YEAR], data from (Federal Reserve Bank of St. Louis., 2023b).

[13] Extensions in Appendix A use data post July 2020.

Note we do not fill or pad next day returns, however where values are missing in lagged sentiments or controls, such as on public holidays, we forward fill using the last preceding value in the series. This is applied to each variable separately. For weekends, lagged values are taken from the last previous trading day, e.g. typically Friday. Finally, each panel includes a control regression consisting of all controls excluding sentiment, to better highlight the additional contribution of sentiment.

To determine if LLM derived sentiments capture, or proxy for different aspects of investor sentiment over different time horizons, we also perform regressions on weekly and monthly frequencies. Weekly models are simple OLS regressions of the same form as Equation (1), and fit next week S&P 500 returns based on lagged sentiments and controls from the prior week. Daily sentiments and controls are re-sampled as follows. For weekly regression models, $Sent_C$ is the average daily sentiment over the prior trading week, Monday to Sunday. We choose to use the average rather than the last value to proxy overall market sentiment of the prior week. The S&P 500 returns R_{t+1} and R_t are calculated as the % returns between the last trading day close of each week, typically Friday to Friday. The remaining controls VIX_t, $Tedrate_t$, $Termspread_t$, and $BAA10Y_t$ are all based on the value of the last trading day of the prior week. We choose to use the values of the last trading day since the VIX can see large fluctuations intra-week and may be driven by short-term daily events, while the remaining controls generally have similar values at this frequency, as they reflect lower frequency macro-economic information.

The monthly regression models differ to the daily and weekly frequency models as we use a different set of controls. We adopt many independent variables from Welch and Goyal (2008) in an effort to control for both macro-level stock-characteristics and interest-rate related aspects, which are relevant at this lower frequency. The OLS model fits the value-weighted continuously compounded returns of the S&P 500 including dividends, $CRSP\_SPvw_{t+1}$, using lagged sentiment and controls, as follows:

$$\begin{aligned} CRSP\_SPvw_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 BW_t + \beta_3 dp_t + \beta_4 dy_t + \beta_5 ep_t + \beta_6 svar_t + \beta_7 bm_t \\ + \beta_8 ntis_t + \beta_9 ltr_t + \beta_{10} tms_t + \beta_{11} dfy_t + \beta_{12} dfr_t + \beta_{13} infl_t + \varepsilon_{t+1}. \end{aligned} \tag{2}$$

Sentiment $Sent_{t,C}$ is the average daily sentiment from each classifier over the prior trading month. We also include the widely cited Baker and Wurgler (2007) Sentiment Index, BW_t, for the prior month. The response variable $CRSP\_SPvw_{t+1}$ and all remaining controls are from Welch and Goyal (2008), with all controls lagged by one month. See Appendix A.6 for variable details.

3.2 Out-Of-Sample Performance Evaluation

Although in-sample analysis is useful for parameter estimation, Welch and Goyal (2008) highlight the discrepancy between in-sample performance and out-of-sample performance. Hence we use several different methods to evaluate both the out-of-sample forecast performance of our sentiment models, and also the economic importance of sentiment when used as part of a trading strategy. Since our study primarily focuses on daily sentiment measures, our out-of-sample evaluations measures are presented for the default daily frequency.

We assess out-of-sample predictive performance via the widely used R^2_{oos} metric of Campbell and Thompson (2008), and the MSPE-adjusted statistic by Clark and West (2007). Equation (1) forecast models are used to calculate $\hat{R}_{t+1}$ based on an expanding window of observations up to t. The resulting forecast errors are then used along with benchmark model forecasts $\bar{R}_{t+1}$ to calculate R^2_{oos} via:

$$R^2_{oos} = 1 - \frac{\sum_{t=1}^{T} \left(R_{t+1} - \hat{R}_{t+1}\right)^2}{\sum_{t=1}^{T} \left(R_{t+1} - \bar{R}_{t+1}\right)^2}. \tag{3}$$

R^2_{oos} is bound $(-\infty, 1]$, and an $R^2_{oos} > 0$ implies the predictive model $\hat{R}_{t+1}$ outperforms the benchmark model $\bar{R}_{t+1}$ in terms of mean-squared prediction error, MSPE. We calculate two sets of R^2_{oos} values based on two different benchmarks. The first benchmark is the control regression, i.e. all controls excluding sentiment. The second is the historical mean of returns. We include this second benchmark as it has been shown to be a stringent out-of-sample benchmark which individual economic variables often fail to outperform, (Welch and Goyal, 2008; Campbell and Thompson, 2008). Both benchmarks forecast each next day return based on observations up to t. Calculating two sets of R^2_{oos} values allows us to better quantify the incremental benefit of sentiment in forecasting returns. In addition, we conduct the right-tailed nested model test of Clark and West (2007) to check if the average MSPE-adjusted statistic of both benchmarks are less than or equal to the predictive model MSPE. A positive test statistic above 1.645 indicates the sentiment model performs better at the 5% significance level.

We also investigate the economic significance of sentiment in predicting returns from an asset allocation perspective. Following Campbell and Thompson (2008), Rapach et al. (2016), Obaid and Pukthuanthong (2022) and others, we compute the certainty equivalent return (CER) gain of a mean-variance investor who allocates between the S&P 500 Index and a risk-free asset each day according to the out-of-sample sentiment model forecast, $\hat{R}_{t+1}$.[14] Each day, the investor allocates w_t in the index, and $1 - w_t$ in the risk free asset, according to:

$$w_t = \frac{1}{\gamma}\frac{\hat{R}_{t+1}}{\hat{\sigma}^2_{t+1}}, \tag{4}$$

where γ is the risk aversion coefficient of three, $\hat{\sigma}^2_{t+1}$ is the variance forecast estimated over an expanding window of historical daily returns, and $\hat{R}_{t+1}$ is the out-of-sample sentiment model forecast. We restrict weights w_t between 0 and 1, and calculate non-cumulative returns each day. By investing according to (4), the investor realizes an average CER of:

$$CER = \hat{\mu} - \frac{1}{2}\gamma\hat{\sigma}^2, \tag{5}$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the average and variance of the portfolio over the evaluation window. The CER gain is then calculated as the difference between the CER of an investor who allocates according to the predictive regression forecast, and the CER of an investor who allocates according to a benchmark portfolio, typically the historical average returns forecast. We calculate both the CER gain of sentiment models against the historical average returns forecast, and the CER gain of sentiment models against the control regression forecast model. This second comparison allows us to directly determine the incremental economic benefit of including sentiment in the returns forecast model (1) when used in a portfolio trading strategy. For completeness, we also calculate the CER gain of the control model against the historical average returns forecast.

In addition to calculating the CER gain for each portfolio over the full sample, we also calculate CER gain values over time. One limitation of reporting CER gain values for the full study period is that this fails to capture if individual model forecasting is sensitive to expanding window size, or if portfolio performance varies over time. For example, if CER gain values of one portfolio only come to dominate later in the sample, this would not be captured. Hence we calculate CER gain values over time for all sentiment models. This allows us to assess the relative performance of the sentiment classifier portfolios to portfolios based on both the control benchmark and the historical

[14] For the daily risk-free rate we use the equivalent daily compounding interest rate as calculated from the daily 3-Month Treasury Bill rate [DTB3], see (Federal Reserve Bank of St. Louis., 2023d).

mean returns benchmark at different points of our study period. To calculate CER gain values over time, we use an initial 3-year estimation period for the purpose of initialising the portfolio allocation model, followed by an initial 3 years of portfolio trading for the purpose of calculating initial CER gain values. Hence CER gains are calculated from 1-1-2006 onwards. We use a minimum of 3 years of portfolio performance to calculate initial CER gain values as we aim to strike a balance between having sufficient data to obtain reasonable initial CER gain values, while also aiming to maintain the majority of data for out-of-sample evaluation. Using these periods has the added benefit of allowing us to capture the particularly noisy period of the GFC, and allows us to observe both ex-ante and ex-post performance of the portfolios. Finally, we report the Sharpe ratios of each sentiment portfolio and the two benchmark portfolios. The Sharpe ratios are calculated from the average portfolio return using sentiment model forecasts net the risk-free rate divided by the standard deviation of the excess portfolio return.

4 Empirical Results

4.1 In-sample Results

In the following section we present in-sample results for daily, weekly, and monthly frequencies over the full dataset. Each panel includes a control regression consisting of lagged controls only, excluding sentiment. We also present several figures illustrating daily sentiment significance over time, based on both expanding and rolling windows of historical observations. All results use robust standard errors.

4.1.1 Returns Forecast - Daily

Table 3 reports daily in-sample results. Consistent with prior literature, and across all classifiers, increases in pessimistic sentiment predict next day return reversals. Almost all sentiments have at least 5% significance, even in the presence of highly significant lagged returns, and the adj. R^2 values of all sentiment models are greater than the control regressions.

The ChatGPT sentiment model performs the best, with the highest adj. R^2 and BIC values, outperforming the TwitterRoBERTa transformer, followed by the traditional dictionary approaches,

Variables	Control	LM	VA	FB	TR	GP
$Sent_t$	-	-0.0203 **	-0.0095 **	-0.0001	-0.0002 **	-0.0016 **
R_t	-0.1155 ***	-0.1311 ***	-0.131 ***	-0.135 ***	-0.1458 ***	-0.1653 ***
VIX_t	0.0067	0.0075	0.0077	0.0070	0.0080	0.0072
$Tedrate_t$	-0.0014	-0.0016	-0.0016	-0.0014	-0.0014	-0.0014
$Termspread_t$	-0.0002	-0.0002	-0.0002	-0.0001	-0.0002	-0.0002
$BAA10Y_t$	-0.0002	-0.0001	-0.0001	-0.0002	-0.0002	-0.0001
Intercept	0.0002	0.0003	-0.0002	0.0003	-0.0001	-0.0003
R-squared	0.0162	0.0176	0.018	0.0171	0.0184	0.0192
Adj. R-squared	0.0152	0.0164	0.0168	0.0159	0.0172	0.0180
AIC	-29,420	-29,425	-29,427	-29,423	-29,429	-29,433
BIC	-29,374	-29,373	-29,375	-29,371	-29,377	-29,381
n	4,958					

Table 3: In-sample daily sentiments and next day market returns.
This table reports regression constants β from model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}$, where R_{t+1} is the next day % change of the S&P 500, and $Sent_{t,C}$ is the overall daily pessimism sentiment measure for classifier C, one of the six sentiment classifiers: Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR) and ChatGPT (GP). R_t is the daily % change of the S&P 500, VIX_t is the daily CBOE Volatility Index/100, and $Tedrate$, $Termspread$ and $BAA10Y$ are measures of daily short, medium, and long-term credit risks, as detailed in Section 3.1. The sample period starts 1st January 2000, and ends 31st July 2020. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

and FinBERT last. While several classifiers have worse, or only slightly higher BIC values relative to the control regression, ChatGPT has the most notable improvement, implying the inclusion of ChatGPT sentiment contributes a marginal amount of new information. In contrast, measures from other sentiment classifiers contribute little or no new information, but rather incorporate information from the control variables. However, the fact all but one of the sentiment measures are statistically significant is worth noting, and these results suggest ChatGPT may be able to better identify more complex information signals from text, beyond that of simpler transformers and dictionaries.

In order to check the controls included are not contributing to or complicating the observed significance of sentiment via suppressor effects, we run a regression model using only sentiment and 1-day lagged returns. Table 4 shows consistently negative and significant sentiment in line with prior results. Overall the adj R^2 values are less, including the control regression, and similar BIC values observed, with ChatGPT performing the best overall.

Finally, in Appendix A.4 we conduct a daily regression model using sentiments from all classifiers, to check if any one sentiment classifier dominates the others head-to-head. The results show no clear winner.

Variables	Control	LM	VA	FB	TR	GP
$Sent_t$	-	-0.0172*	-0.0078**	-0.0001	-0.0002**	-0.0016**
R_t	-0.1203***	-0.134***	-0.1337***	-0.1396***	-0.1497***	-0.1694***
Intercept	0.0002	0.0004**	0.0000	0.0004**	0.0001	0.0000
R-squared	0.0144	0.0154	0.0156	0.0153	0.0164	0.0173
Adj. R-squred	0.0142	0.015	0.0152	0.0149	0.016	0.0169
AIC	-29,419	-29,422	-29,423	-29,421	-29,427	-29,431
BIC	-29,399	-29,396	-29,397	-29,395	-29,401	-29,405
n	4,958					

Table 4: In-sample daily sentiments - suppressor effects check.
In order to check that suppressor effects are not confounding results, this table reports regression constants β from model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \varepsilon_{t+1}$ for all five classifiers. Results show consistently negative and significant sentiment coefficients. The sample period starts 1^{st} January 2000, and ends 31^{st} July 2020. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

4.1.2 Returns Forecast - Weekly

Variables	Control	LM	VA	FB	TR	GP
$Sent_t$	-	-0.1375 *	-0.0270	-0.0004	-0.0007	0.001
R_t	-0.0728	-0.1036 *	-0.0871	-0.0861	-0.0979	-0.0645
VIX_t	0.0002	0.0003	0.0002	0.0002	0.0003	0.0002
$Tedrate_t$	-0.0089	-0.0100 *	-0.0095 *	-0.0089	-0.0088	-0.0089
$Termspread_t$	-0.0014	-0.0018 *	-0.0015	-0.0014	-0.0016	-0.0014
$BAA10Y_t$	0.0012	0.0015	0.0014	0.0012	0.0012	0.0012
Intercept	-0.0002	0.0007	-0.0013	0.0002	-0.0012	0.0001
R-squared	0.0195	0.0235	0.0204	0.0199	0.0210	0.0196
Adj. R-squared	0.0149	0.0179	0.0149	0.0144	0.0154	0.0140
AIC	-4,827	-4,829	-4,826	-4,826	-4,827	-4,825
BIC	-4,792	-4,790	-4,786	-4,786	-4,787	-4,785
n	1,072					

Table 5: In-sample weekly sentiments and next week market returns.
This table reports regression constants β from model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}$, where R_{t+1} is the % change of the S&P 500 for the following week, $Sent_{t,C}$ is the average daily pessimism sentiment measure for classifier C for the prior week, R_t is the % change of the S&P 500 for the prior week, VIX_t is the CBOE Volatility Index/100 as of market close of the prior week, and $Tedrate$, $Termspread$ and $BAA10Y$ are measures of daily short, medium, and long-term credit risks, taken at market close of the prior week, as detailed in Section 3.1. The sample period starts 1^{st} January 2000 and ends 31^{st} July 2020. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5 reports in-sample results of average weekly sentiment models. At this frequency, results suggest the inclusion of sentiment leads to little model improvement. All BIC values are lower than the control model, and only LM and TR have imrpoved adj. R^2. Lagged returns are also less significant at this frequency compared to daily results. While all sentiment coefficients are negative, except for ChatGPT, only LM is significant at 10%. The fact the coefficient of ChatGPT is the lowest, is positive, and has the lowest adj. R^2 for weekly data, suggests ChatGPT may be capturing information at a higher frequency relative to the other classifiers. This is further reinforced by the fact that for daily data, the inclusion of ChatGPT sentiment yields the best performance. These results support the work of Ke et al., 2019, that news information is quickly absorbed into prices, and suggests that daily news summaries fail to consistently capture persistent investor sentiment at lower frequencies. For completeness, we present a weekly regression model with sentiments from all classifiers in Appendix A.4, where some significance of ChatGPT is observed. However, in this model LM dominates, and the fact the signs are opposite suggest this may be in part due to collinearity.

4.1.3 Returns Forecast - Monthly

Variables	Control_1	Control_2	LM	VA	FB	TR	GP
$Sent_t$	-	-	0.482	0.4297 ***	0.0065	0.0043	0.0004
BW_t	-	-0.0044	-0.0048	-0.0059	-0.0076	-0.0071	-0.0051
dp_t	0.0114	0.0152	-0.0389	-0.07	-0.0533	-0.041	0.0084
dy_t	0.1007	0.0866	0.1284	0.1481 *	0.1601	0.1424	0.0885
ep_t	-0.0021	-0.0023	-0.0049	-0.0116	-0.0034	-0.0031	-0.0031
$svar_t$	0.222	0.1952	0.2514	0.2265	0.2301	0.1864	0.2127
bm_t	-0.0214	-0.0284	-0.0017	0.0177	-0.039	-0.0398	-0.0259
$ntis_t$	0.71 **	0.6967 **	0.6115 *	0.4648	0.6823 **	0.6742 *	0.6882 **
ltr_t	0.0326	0.0233	0.0279	0.016	0.0071	0.0053	0.0206
tms_t	-0.4896 *	-0.5121 *	-0.4873 *	-0.5814 **	-0.5643 **	-0.4764	-0.5196 *
dfy_t	-2.0185	-1.9422	-2.0085	-2.2045	-2.5022	-2.3945	-1.9385
dfr_t	1.3593 *	1.613 **	1.0812	0.5545	1.664 **	1.7227 **	1.5972 **
$infl_t$	0.889	0.8961	0.7345	0.5982	0.685	0.7082	0.849
Intercept	0.4701 **	0.4272 *	0.3633	0.317	0.4438 *	0.4329 *	0.4052 *
R-squared	0.1104	0.112	0.1134	0.135	0.1174	0.1135	0.1085
Adj. R-squared	0.0687	0.0665	0.0638	0.0865	0.068	0.0639	0.0585
AIC	-854.4	-852.9	-848.3	-854.4	-849.4	-848.4	-846.9
BIC	-808.7	-803.7	-795.8	-801.8	-796.9	-795.8	-794.4
n				247			

Table 6: In-sample monthly sentiments and next month market returns.
This table reports regression constants β for monthly regression model Equation (2), which fits next month S&P 500 returns, $CRSP\_SPvw_{t+1}$, from prior month text sentiment and controls. We include controls from Welch and Goyal (2008), which are better suited to capturing monthly frequency effects than daily and weekly model controls. Two measures of sentiment are included; $Sent_t$ is the average monthly sentiment derived from news summary text, and BW_t is the well-known Baker and Wurgler (2007) sentiment index. Control variable details are provided in Appendix A.6. Two control models are included for clarity. The first is comprised of all controls excluding both text-based sentiment and the Baker Wurgler sentiment index, while the second excludes only text-based sentiment. The sample period starts 1^{st} January 2000 and ends 31^{st} July 2020. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Monthly in-sample results are reported in Table 6. Similar to weekly frequency, averaged monthly text sentiments from daily news summaries are not significant, with the exception of VA. Interestingly at this frequency all coefficients are positive, a notable difference to the negative coefficients observed at higher frequencies. Overall, results suggest the inclusion of sentiment does not improve model performance, with the exception of VA, which sees an improved adj. R^2 value over both control models. The significance of VA sentiment is unexpected, however the lower BIC value relative to both controls, combined with the substantially lower and non-significant coefficients for $ntis$ and dfr values relative to all other models, suggests that $Sent$ is capturing the explanatory power from these two variable. Further, Baker Wurgler sentiment is negative, but not significant, and there is no improvement between control models 1 and 2. This supports prior literature that Baker and Wurgler (2007) sentiment is a contemporaneous explainer rather than a predictor of future returns.

It is worth noting there are less observations at this lower monthly frequency than both daily and weekly models. Combined with the increased number of independent variables, the reported significance of aggregated VA daily news sentiments should be viewed cautiously, as reduced sample size in combination with the greater number of control variables increases the chance of overfitting. For completeness, we present monthly models with sentiments from all classifiers in Appendix A.4. In this model specification ChatGPT sentiment is negative and statistically significant at 5%, however since VA is significant and positive, this is likely due to collinearity with one sentiment measure counterbalancing the other.

4.1.4 Sentiment Significance Over Time

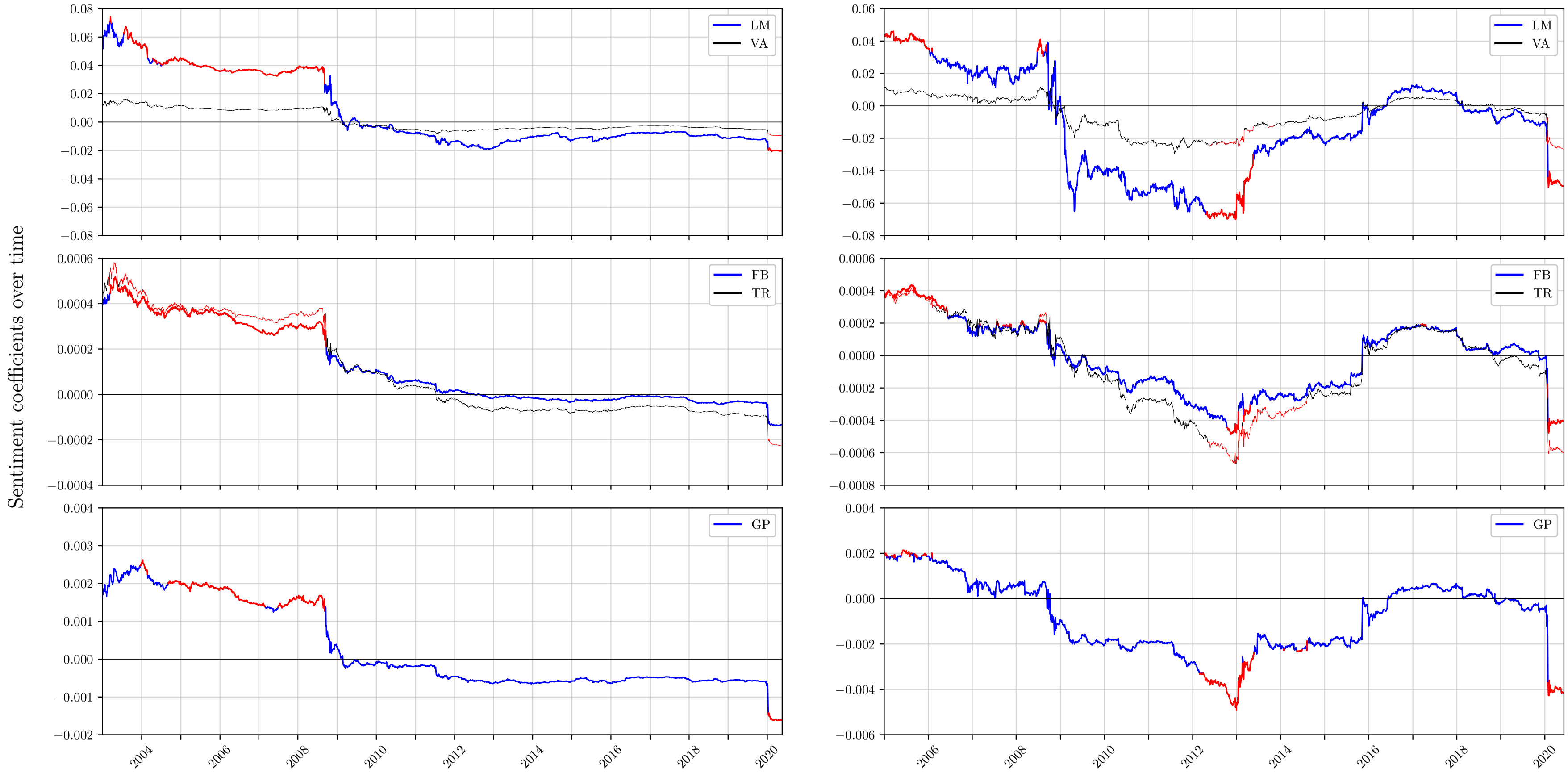
Both Garcia (2013) and Tetlock (2007) show the role of sentiment in forecasting returns can vary over time, such as during periods of market uncertainty. While useful, the in-sample results of Table 3 do not capture these potential temporal differences. Hence we use Equation (1) to evaluate sentiment coefficients over time via two approaches.

The first is based on an expanding window of daily observations. Starting with an initial 3-year estimation period, the sentiment coefficients of each classifier are progressively calculated based on data up to each point t, and plotted over time. While the choice of an initial 3-year estimation period is somewhat arbitrary, we choose this as it is a reasonable trade-off for obtaining initial model estimates, while retaining the majority of data for evaluation. This choice also allows us to capture sentiment changes over the particularly noisy period of the Global Financial Crisis (GFC). The second method calculates sentiment coefficients over time using a 5-year rolling window approach.

Since expanding window results suggest sentiments coefficients smooth out after roughly 4 years, we use a 5-year rolling window to reduce noise. Hence out-of-sample predictions start from 2005 onwards. Results are presented in Figure 3. Note the last value of each series for the expanding window approach corresponds to the full in-sample daily coefficients of Table 3.

Figure 3 shows that while sentiment varies over time, ChatGPT sentiments closely follow those of traditional dictionary and simpler transformer models. While initially sentiment coefficients pre-GFC are all positive and significant for all classifiers at the 5% level, sentiment coefficients sharply drop to negative with the onset of the GFC. Sentiment coefficient values remain negative for the rest of the sample, however for the expanding window approach are not significant until the onset of COVID-19. The rolling 5-year window results show similar behaviour. Initially sentiment values are positive, with periods of significance, however turn sharply negative around the GFC. Values then remain negative for a long time, with sentiment significant for a period of time at the tail end of the European Debt Crisis around the mid 2010's. Sentiment values then slowly increase for all classifiers towards the end of sample, while significance vanishes until the onset of COVID-19, when values sharply turn negative and significant once again.

The results highlight that ChatGPT classifies sentiment in a similar manner to dictionary and transformer methods, with sentiment negative for the majority of the sample. While sentiment significance varies over time, it is primarily clustered around specific periods, such as COVID-19, and these periods are consistent across classifiers. Since simple dictionary methods cannot account for semantic context, results suggests that sentiment information is at least partly transmitted through simple unique features of the news text summaries, specifically through the negative words present. However, since ChatGPT and TwitterRoBERTa dominate simple dictionaries in in-sample performance per Table 3, then assuming behavioural models of next-day return reversals hold, the results also suggest ChatGPT and TwitterRoBERTa are identifying information through semantic context, which is a contributing factor to outperformance.



Panel A: U.S. Stock News Expanding Window

Panel B: U.S. Stock News 5-Year Rolling Window

Figure 3: Sentiment Coefficients over Expanding and Rolling Windows from U.S. Stock News Summary Text.
These graphs show sentiment coefficient values β_1 of sentiment models Equation (1), calculated over both an expanding and 5-year rolling window of daily historical observations, from 1^{st} January 2000 up to time t. The first three and five years are used to calculate initial values of expanding and rolling results respectively. The five sentiment measures are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), and ChatGPT (GP). Classifiers are grouped in pairs; dictionaries, transformers, and ChatGPT. Values are highlighted in red if significant at time t, for $p < 0.05$ based on robust standard errors.

4.2 Out-of-sample Results

4.2.1 Out-of-sample R^2, CER, Sharpe Ratios, and CER over time

Our main out-of-sample results are detailed in Table 7. We present the out-of-sample R^2 performance (R^2_{OOS}) of next day sentiment model forecasts against two benchmarks; the control model consisting of all controls excluding sentiment, and the historical mean of returns, i.e. constant returns model. We report CER and CER gain values of portfolios which trade based on next day sentiment model forecasts. To highlight the added contribution of sentiment, we present CER gains over two benchmarks, the control model portfolio, and the historical mean returns portfolio. Finally, we present Sharpe ratios of each sentiment model, the control model, and mean historical returns model. See Section 3.2 for detailed methodology.

	U.S. Daily Stock News Summaries 1-1-2000 to 31-7-2020									
	R^2_{OOS} (%)		CER gain (%)			CER (%)		Sharpe ratios		
Classifier	$Sent._{Cont.}$	$Sent._{Hist.}$	$Sent._{Hist.}$	$Sent._{Cont.}$	$Cont._{Hist.}$	Sent.	Cont.	Sent.	Cont.	Hist.
LM	0.02*	0.78***	5.786	1.36		7.349		0.706		
VA	0.11**	0.86***	4.683	0.257		6.246		0.621		
FB	-0.06	0.70***	6.496	2.07	4.427	8.059	5.989	0.77	0.597	0.247
TR	0.09*	0.85***	**7.188**	**2.761**		**8.75**		**0.824**		
GP	**0.22****	**0.97*****	6.042	1.615		7.604		0.728		

Table 7: Out-of-sample analysis results.
Summary out-of-sample results for daily sentiment models which forecast next day % returns of the S&P 500 index from an expanding window of observations, according to Equation (1). The five sentiment models are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), and ChatGPT (GP). Two sets of R^2_{oos} are presented; sentiment model forecasts vs control model forecasts, and sentiment model forecasts vs historical mean returns model forecasts. Significance of the right-tailed nested model test of Clark and West (2007) is shown for $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.[15] Certainty equivalent return (CER) % and CER gain % values of portfolios based on both sentiment model forecasts and benchmark model forecasts are presented, calculated per Section 3.2 and annualized.[16] Three sets of CER gain values are presented for clarity; CER gain of sentiment portfolios over historical mean returns portfolio, CER gain of sentiment portfolios over the control portfolio, and the CER gain of the control portfolio over the historical mean returns portfolio. Raw CER values for both sentiment and control portfolios are shown. Annualized Sharpe ratios for sentiment portfolios, and both benchmark portfolios are presented.[17] The sample period starts 1st January 2000 and ends 31st July 2020. R^2_{oos} is calculated from all data 1st January 2003 onwards, while CER, CER gain, and Sharpe ratios are calculated from 1st January 2006 onwards.

Across all evaluation metrics, ChatGPT and TwitterRoBERTA sentiment models outperform all others, followed by FinBERT and dictionary-based sentiment models last. All sentiment models have positive R^2_{OOS} values when benchmarked against both the control and historical mean returns, except for FinBERT, and most report significance for the Clark and West nested model test,

[15] Clark and West equal predictive performance test of nested models. We use the right-tailed test, which tests the null hypothesis of equal model performance against the alternative that the larger sentiment model outperforms the nested model in terms of MSPE.

[16] CER and CER gain values are annualized by multiplying by 250.

[17] Sharpe ratios are calculated from daily portfolio performance per Section 3.2. The daily risk-free rate is calculated as the daily compounding equivalent of the 3-Month Treasury Bill rate [DTB3], see (Federal Reserve Bank of St. Louis., 2023d). Since the resulting Sharpe ratios are daily, we annualize by multiplying by $\sqrt{250}$.

implying the inclusion of sentiment improves model MSPE accuracy and performance. ChatGPT has the largest R^2_{OOS} values, followed by TR, while FB and LM have the lowest.

All sentiment portfolios have notably larger CER gain % when benchmarked against both the control portfolio and the historical mean returns portfolio. Here, ChatGPT has the third largest CER gain %, outperformed by both TR and FB, with dictionary methods performing last, with both dictionary portfolios LM and VA reporting lower CER and CER gain values relative to other portfolios. FB sees somewhat mixed results, with the lowest R^2_{OOS} values close to zero when benchmarked against the control model, yet positive CER gain relative to the control portfolio benchmark, beating ChatGPT, LM and VA. These mixed results are most likely a result of portfolio weights being restricted to between [0,1], preventing reduced CER gain performance from future incorrect short positions. Also, several models fail to reject the Clark and West null hypothesis at the 5% level when benchmarked against the control model, hence their performance should be viewed with caution.

The annualized Sharpe ratios of all sentiment portfolios are greater than both the control and historical means portfolios, with ChatGPTs portfolio performing in the middle of the pack, outperforming both dictionary methods. Together, the R^2_{OOS}, CER gain %, and Sharpe ratio results point to improved model forecasting and portfolio performance when sentiment is taken into account, and suggest ChatGPT and transformer models offers improved performance compared to dictionary methods.

Overall, our results are in agreement with previous literature of (Obaid and Pukthuanthong, 2022; Birru and Young, 2022), and others, noting some slight differences. First, our forecast period is larger than some studies, with our model forecasts starting from 1^{st} January 2003 onwards. Hence, our out-of-sample portfolio assessment covers the particularly noisy period of the GFC, which some other studies fail to capture. Second, our proxies for investor sentiment are single daily news summaries, rather than aggregated sentiment over several news sources. Nevertheless, we proxy sentiment from text, and it is likely sentiment signals conveyed in text are stronger than signals transmitted via other mediums. For example, results from Obaid and Pukthuanthong (2022) suggest sentiment signals from news article photos are roughly four times weaker than sentiment signals conveyed in news article texts. Coincidentally, our reported R^2_{OOS} and CER values are roughly four times larger than their results. Even though we use a slightly different set of controls to prior studies, the fact we obtain similar and complimentary results lends support to our methodology and findings.

4.2.2 CER over time

Although useful for out-of-sample evaluation, one limitation of CER gain results of Table 7 is that they do not capture portfolio performance over time. Hence if one portfolio only outperforms the others towards the end of the sample period, this would not be identified. Therefore we present CER gain values over time in Figure 4 to further clarify sentiment model forecasts and portfolio performance.

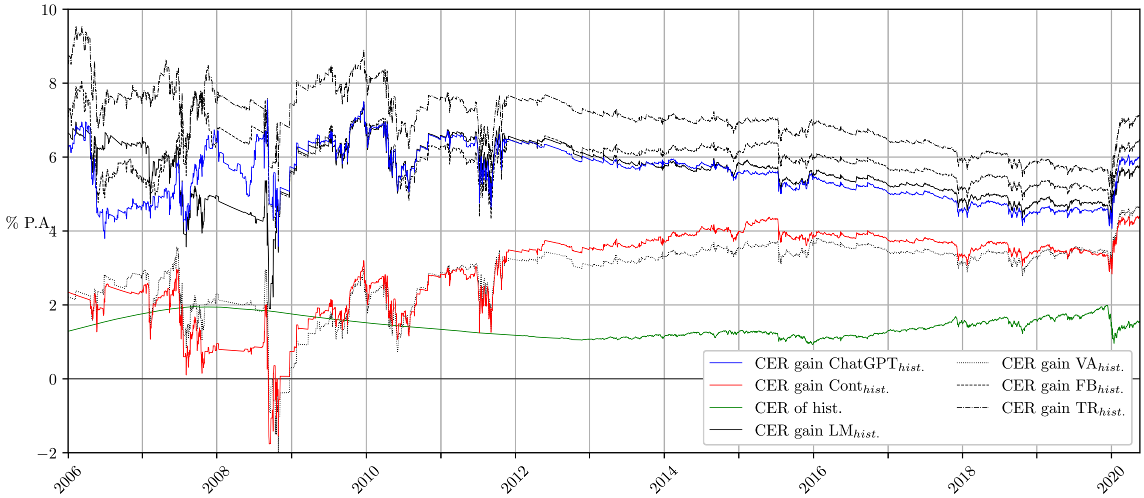


Figure 4: CER gain values over time.
These graphs show annualized changes in CER gain values of daily sentiment portfolios vs the historical mean returns portfolio over time. The two best performers are TwitterRoBERTa and FinBERT transformers, followed closely by ChatGPT (blue). Daily portfolio allocations are made between the S&P 500 Index and the daily risk-free rate based on next day model forecasts over an expanding window of observations. An initial 3-year estimation period is used for initial model forecasts and initial next day portfolio allocations, (4). CER gain values are then calculated based on an expanding window of daily portfolio returns, with 3 years of initial trading used to calculate initial CER values, (5). Hence CER gain values are presented from 1^{st} January 2006 onwards, with values at time t representing CER gain of portfolios trading from 1^{st} January 2003 to t. For completeness, we include the CER gain of the control portfolio above the historical mean returns model (red), and the CER of the historical mean returns model (green). Note the last value of each series correspond to full sample results of Table 7.

Results show sentiment-based portfolios consistently outperform the historical mean returns benchmark over the full sample. The ChatGPT portfolio outperforms several alternatives for the majority of the sample, and importantly, the CER gains of all sentiment portfolios exceed those of the control portfolio, although VA underperforms at several points. Overall, results demonstrate the inclusion of sentiment leads to better portfolio performance.

In the lead-up to the GFC there is initial variability and decline in CER gain values, and the collapse of Lehman Brothers in late 2008 coincides with a sharp drop across all portfolios. All CER gain values reach their lowest levels at this time, with several briefly turning negative, in comparison to the robustness of the mean of historical returns. Values for all sentiment portfolios then recover over the next 12 months, followed by a noisy period from 2010 through 2012, as concerns of a European Debt Crisis grow. Following 2012, CER gains remain fairly consistent and stable until the onset of COVID-19 in 2020. At this time, a roughly 50% drop in the CER of the historical mean return portfolio, i.e. the benchmark (green), results in artificially elevated CER gain values for all portfolios, including the control portfolio (red). Note that all CER gains trend slightly downward post-2012, in part due to the slight uptrend in the CER of historical returns, a reflection of the robustness of the historical means returns model, consistent with Campbell and Thompson (2008).

These results show the inclusion of sentiment leads to significant CER gains above those of the control portfolio (red), with ChatGPT-derived sentiments providing robust outperformance over time, outperforming simple dictionary methods. While ChatGPT does not strictly outperform the FinBERT and TwitterRoBERTa transformers in terms of CER gain, its performance is only marginally less, and remains a top contender for the majority of the sample.

5 Underlying Economic Channels

Our results present evidence of daily news sentiment predictability of next-day stock market returns. In this section we investigate various potential economic channels that may underpin these findings. Prior literature suggests mispricing occurs most prominently when fundamentals are more uncertain, and limits to arbitrage are higher, when the effects of investor irrationality are exacerbated. Hence we investigate sentiment significance during different conditions of uncertainty according to several indicators from prior literature, including; economic uncertainty, economic policy uncertainty, times of recession, and overall business conditions. For simplicity, we follow a similar approach to Birru and Young (2022), which involves first separating daily observations into periods of lower (higher) uncertainty, or better (worse) conditions, and then conducting daily in-sample regressions for each group separately using the same daily model specification of (1). See Sections 5.1 to 5.4 for details. Summary results are presented in Table 8.

Across all conditions tested, results demonstrate findings consistent with prior literature of heightened sentiment induced mispricing during more uncertain times. We observe negative sentiment coefficients across all conditions, consistent with prior results, with statistically significant sentiments predicting daily aggregate market returns. Further, the inclusion of sentiment from ChatGPT results in greater improvements to adj. R^2 and BIC values relative to sentiments from simpler transformers and dictionary methods, suggesting ChatGPT sentiments may better capture sentiment induced mispricing effects, which are particularly notable during times of greater uncertainty and reduced economic activity.

		Better conditions						Worse conditions					
		Control	LM	VA	FB	TR	GP	Control	LM	VA	FB	TR	GP
FEARS	$Sent_t$	-	-0.0617	-0.0078	-0.0003	-0.0004	-0.0029	-	-0.1318 *	-0.0509 **	-0.0004	-0.0006	-0.0048 *
	Adj. R-sq	0.071	0.073	0.069	0.071	0.073	0.077	0.068	0.084	0.083	0.070	0.071	0.081
	BIC	-2118	-2113	-2112	-2113	-2114	-2115	-1975	-1976	-1976	-1970	-1971	-1975
	n				378						378		
Economic Uncertainty	$Sent_t$	-	0.0083	0.003	0.0001	0.0001	0.0005	-	-0.0772 ***	-0.029 ***	-0.0008 ***	-0.0011 ***	-0.0073 ***
	Adj. R-sq	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.037	0.045	0.045	0.050	0.061	0.065
	BIC	-6528	-6521	-6521	-6522	-6522	-6522	-5192	-5194	-5194	-5200	-5211	-5216
	n				952						1,012		
NBER Recession	$Sent_t$	-	-0.0079	-0.0032	0	0	0.0001	-	-0.0351	-0.0374	-0.0003	-0.0008	-0.0063 **
	Adj. R-sq	0.002	0.002	0.002	0.002	0.002	0.002	0.044	0.043	0.047	0.043	0.046	0.057
	BIC	-28007	-28000	-28000	-27998	-27999	-27998	-2538	-2532	-2534	-2532	-2534	-2540
	n				4,401						559		
ADS Business Conditions Index	$Sent_t$	-	-0.0269 *	-0.0111 **	-0.0002 *	-0.0003 **	-0.0013	-	-0.0374	-0.0156	-0.0001	-0.0003	-0.0037 **
	Adj. R-sq	0.020	0.023	0.023	0.023	0.025	0.022	0.036	0.036	0.036	0.035	0.036	0.042
	BIC	-6356	-6353	-6353	-6353	-6355	-6352	-4930	-4925	-4925	-4924	-4925	-4931
	n				992						992		

Table 8: Sentiment during different market conditions.
This table summarises daily sentiment significance for Equation (1) over several different conditions. We test periods of; elevated societal pessimism (Da et al., 2015), periods of lower and higher economic uncertainty (Baker et al., 2016), periods of recession (Federal Reserve Bank of St. Louis., 2023c), and overall business conditions (Federal Reserve Bank of Philadelphia., 2023). For brevity, sentiment coefficients, adj. R^2 and BIC values are presented. Across all conditions, sentiment is more negative, significant, and yields higher adj. R^2 and BIC values when societal concerns and economic uncertainty is higher, and when business conditions are worse. ChatGPT consistently dominates all other sentiment classifier models when conditions are worse. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

5.1 Economic uncertainty

Da et al. (2015) suggests that when uncertainty is greater, the effects of investor irrationality and sentiment-induced mispricing may be greater. To investigate uncertainty as a potential channel, we use their Fears30 Index to define periods of lower and higher uncertainty based on days corresponding to the lowest (highest) quintiles. One unique aspect of this index is that it proxies societal uncertainty from Google search terms, and hence is a reasonable proxy for overall societal beliefs about uncertainty which are not strictly based on fundamentals. In sample regressions are then conducted for both groups of lowest and highest uncertainty, and results presented in Table 8.[18]

[18] Since the FEARS index spans 1/7/2004 to 30/12/2011 we define low and high quintiles from days within this period only. We use Fears30, however the two alternatives make almost no difference to the results.

Results show sentiment is only significant during periods of elevated economic concern. During these times, the control regression has lower adj. R^2 and BIC values, which is expected during more uncertain times. Further, when FEARS is greatest, all sentiment coefficients are larger, and the inclusion of sentiment corresponds to greater adj. R^2 values and mostly constant BIC values. Notably, while dictionary methods perform the best, they are closely followed by ChatGPT. These results are consistent with temporary sentiment induced mispricing, and since these sentiments are based on end-of-day summaries, supports the second view of Tetlock (2007), that media influences investors attitudes and expectations.

5.2 Economic Policy Uncertainty

Shocks to economic policy have been linked to increased stock market volatility, and reduced investment and employment, as well as reduced economic performance. We investigate this potential channel by identifying days corresponding to the lowest (highest) quintiles of economic policy uncertainty index of Baker et al. (2016), and then running in-sample regressions for both groups.[19]

Results show sentiments are negative and highly significant during periods of greater policy uncertainty, but almost zero and not significant during periods of lower uncertainty. The inclusion of ChatGPT sentiments results in the largest improvement in both adj. R^2 and BIC values, followed by transformers and dictionaries last. Note the low adj. R^2 values are almost negligible for periods of low policy uncertainty, however this is likely due to the grouping of daily observations from the monthly frequency economic policy uncertainty index. Note while the adj. R^2 values are slightly negative when uncertainty is lower, the unadjusted R^2 values are positive for each model.[20]

5.3 Recession Periods

Garcia (2013) demonstrates the predictability of stock returns from news-based sentiments is higher during times of recession. We investigate this channel by using the daily NBER recession indicator to sort our sample into two groups, periods of recession and non-recession, and then perform

[19] As the economic policy uncertainty index is monthly, we define low (high) daily observations based on the months corresponding to the lowest (highest) uncertainty quintiles. The monthly-based quintile with the highest uncertainty has roughly an extra two months of trading days, hence the different number of observations.

[20] R^2 values are between 0.0046 for the control, to 0.0053 for ChatGPT. The negative adj. R^2 values are likely partly due to small sample size.

regressions for each.[21]

Results show during non-recession times the inclusion of sentiment yield no improvement to adj. R^2 or BIC values. In contrast, during recession times ChatGPT sentiment is significant, and the inclusion of this sentiment leads to significantly improved adj. R^2 and BIC values compared to other classifiers. While sentiments from other classifiers are not significant during recession times, this may be due to the limited number of recession days over our 20 year period. The fact ChatGPT sentiment is significant could suggest it is better at identifying sentiment edge-cases, such as particularly subtle text signals. Nevertheless, results for ChatGPT are in agreement with prior literature and suggest ChatGPT may better capture pessimism during recessions, Garcia (2013).

5.4 Economic Activity

Finally, we investigate if return predictability differs during periods of better and worse economic activity using the lowest and highest quintiles of the Aruoba-Diebold-Scotti (ADS) Business Conditions Index.[22] This index aims to track business conditions at a daily frequency, and hence identifies a different subset of data compared to times of recession or uncertainty.

Compared to the other economic channels investigated, the results are somewhat mixed. For days corresponding to the quintile of better economic conditions, sentiments from all classifiers, excluding ChatGPT, are significant. Conversely, when business conditions are at their worst, only ChatGPT sentiment is significant. Also, the inclusion of ChatGPT sentiment when business conditions are worse is the only case when both adj. R^2 and BIC values improve. While it is unexpected that other sentiment measures are not (are) significant during worse (better) conditions, this may be caused by small sample size. Still, all sentiment coefficients are negative and all result in either no change, or marginally improved adj. R^2 values.

[21] Roughly 19 months are classed as 'recession' over our study period. SeeFederal Reserve Bank of St. Louis. (2023c)

[22] Federal Reserve Bank of Philadelphia. (2023). *Aruoba-Diebold-Scotti Business Conditions Index [ADS]*. Federal Reserve Bank of Philadelphia. https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/ads. Accessed: 1 August 2023.

6 LLM Text Recall Sentiment

In an extension of our main results, we investigate both ChatGPT and BARDs ability to recall historical information over time.[23] The approach involves first separately prompting each LLM to recall the most important historical financial information for each day over our 20 year study period, and collating the responses into two text universes.[24] Then, similar to Section 2 we separately prompt both LLMs to classify each daily recalled text summary. In addition to cross-classifying ChatGPT text using BARD and vice versa, we also classify both text universes using traditional dictionaries and transformers; LM, VA, FB, and TR. Finally, we follow the same in-sample and out-of-sample evaluation procedure detailed in Section 3, with the goal of evaluating if LLM recalled text and sentiments differ substantially from daily historical news text summaries. Results and further details are provided in Appendix A.

In summary, we observe very similar results for sentiments derived from LLM-recalled text, with sentiments negative and significant in-sample, with similar coefficients at the daily frequency, Table 9. We also observe similar out-of-sample results Table 12, with very similar R^2_{OOS} and CER gain values observed for both text universes, with ChatGPT performing the best, followed by BARD, transformers, and dictionaries last. Figure 7 shows CER and CER gain values over time are also very similar, with ChatGPT outperforming for much of the sample. Similar patterns over time are also observed independent of classifier, including shocks to CER values during the GFC, noise around the European Debt Crisis, and a sharp spike at the onset of COVID-19 in early 2020. These results suggest ChatGPT and BARD recalled text closely mirrors the main themes and sentiments conveyed in summary news text over time.

There are some notable differences compared to news-based sentiments however. In particular, weekly and monthly results, Tables 10 and 11 show LLM-text as significant in several instances. Weekly results show all sentiment models have improved adj. R^2 and BIC values, with ChatGPT and BARD outperforming the rest, while monthly BARD text shows all sentiments significant.

[23] While other LLMs exist, such as Meta's LLAMA (Touvron et al., 2023), and Bloomberg's BloombergGPT (Wu et al., 2023), we focus on ChatGPT and BARD as they are currently the most widely used. While we would like to investigate BlooombergGPT, as it has been trained on a financial-specific corpus, it is currently not available for external use. We do not focus on non-english LLMs such as Baidu's Ernie as they are primarily fine-tuned on non-English tasks. (Baidu Research. [2023]. *Introducing ERNIE 3.5: Baidu's Knowledge-Enhanced Foundation Model Takes a Giant Leap Forward*. Baidu Research. http://research.baidu.com/Blog/index-view?id=185. Accessed: 27 June 2023) Further, since OpenAI/Microsoft and Google are the current leaders in LLM development, and since their resources are significantly greater than most other western developers, it is likely ChatGPT, BARD, and their future iterations will continue to dominate this space (Gudibande et al., 2023).

[24] Using LLM-generated texts may introduce endogeneity issues, such as knowledge of future events. However we attempt to mitigate this by directing both LLMs to focus on the most important contemporaneous information in an attempt to avoid lookahead or same-day endogeneity effects. We also perform cross-classification of both texts, i.e. ChatGPT classifies BARD text and vice versa, alongside dictionaries and simpler transformers, and by using simple dictionaries we avoid potential future information let slip via contextual reference. Also, prior news-based results allow for a direct comparison to LLM recalled text sentiments, and can help shed light on potential LLM recall bias.

These in-sample result differences at these lower frequencies likely point to LLMs recalling information relevant to future events and look-ahead bias, as specific events may have information content across both higher and lower frequencies. Both expanding and rolling window results support this, with similar overall patterns observed between news text sentiments, and LLM text universe sentiments over time. Sentiment coefficients are consistently much larger and significant for LLM text, and sentiment is notably more negative and significant for longer for LLM text universes post-GFC.[25] These similarities suggest LLMs are recalling text that contain not only particularly strong negative signals, since simple dictionaries are more significant, but text that contains information relevant at lower frequencies. In addition, monthly results for BARD-recalled text suggest that sentiment is more significant across daily, weekly, and particularly monthly frequencies. Assuming behavioural theories of investor overreaction and market reversals hold, this could indicate BARD may be recalling text that contains more information, or better reflects market sentiments over multiple frequencies than ChatGPT. [26]

The results of LLM-recalled text highlight the potential use of LLMs to aid in the summarisation of important financial information. For example, LLMs could be used by retail investors to summarise key topics and themes from various news sources, or be used as a 'co-pilot', helping practitioners to compare news coverage, or help to identify specific financial information. Such implementation could aid with time-constraints and 'attention' issues, (Da et al., 2011). However potential users should keep in mind they will likely be fully reliant on developer disclosure as to model details and potential limitations and biases, such as those resulting from moderation. [27] The tendency to give more politically correct and diplomatic answers may be detrimental to accuracy, particularly when dealing with sensitive topics. The recent paper by Chen et al. (2023) provides evidence that although ChatGPT responses are becoming 'safer', performance may be stagnating, and even be regressing. ChatGPT has also been shown to have a significant left-leaning slant, and is biased against right-winged views (Motoki et al., 2023; Rozado, 2023; McGee, 2023; Baum and Villasenor, 2023). LLMs could hold other biases, such as against particular individuals, or compa-

[25] See both expanding and rolling window results, Figures 3, 5, and 6.

[26] Interestingly, ChatGPT notably outperforms BARD for the BARD text universe monthly frequency in terms of BIC magnitude, and in addition has the highest adj. R^2 value overall, suggesting self-classification bias may be less of an issue. Still, the results suggest a fundamental difference between news summary text and LLM recalled text. Note an important difference between ChatGPT and BARD is that BARD is being continuously updated live, and is connected to the internet, while the ChatGPT model used is only trained on data up to September 2021. This access to a fundamental 'source of truth' may help explain the consistently higher outperformance for all classifiers for the BARD text universe, in particular the out-of-sample results. However it remains unclear why ChatGPT outperforms BARD for the BARD text universe.

[27] There is incentive for developers to keep their intellectual property private, and while there are currently known differences between ChatGPT and BARD, this may not always be the case. Since both ChatGPT and BARD are closed models, there is no way to know what datasets they have been exposed to, trained on, or if they have already been exposed to an evaluation dataset corresponding to a given prompt text. See 'data leakage' (Aiyappa et al., 2023). While the goal of moderation is to avoid harmful responses and prevent exploits such as jailbreak attacks (Li et al., 2023), it can also result in factually incorrect answers and censoring of information.

nies, and such biases could have detrimental impacts.[28] Hence clarity of the moderation process, and potential biases of both LLMs and their developers is an important part of any implementation.

7 Conclusion

In this study we use ChatGPT to classify sentiments from daily end-of-day U.S. Stock News text summaries, and use these sentiments to forecast future returns of the S&P 500 Index. We make three contributions to the literature. First, we demonstrate ChatGPT can classify summary market-level financial text from the perspective of a financial analyst. Second, we show these sentiments proxy for aggregate investor sentiment and forecast future return reversals of the S&P 500 Index, consistent with prior behavioural model literature of investor overreaction (De Long et al., 1990; Campbell et al., 1993). Finally, we provide evidence that ChatGPT and BARD can recall daily news summaries from the perspective of a financial analyst, and that incorporating ChatGPT-derived sentiments from LLM-recalled text leads to superior economic performance compared to portfolios that incorporate sentiments from BARD, simpler transformer models, and traditional dictionary approaches.

LLMs have potentially superior contextual information processing of specific topics or themes beyond that of simpler transformer models and context-indifferent word frequency methods. This greater context awareness leads to better identification of aggregate market sentiment, and superior short-term economic performance when taken into account. Further, results suggest LLMs can identify different aspects of sentiment from text, such as information on different frequencies, and the presence of persistent effects.

Our study serves to highlight the versatility and applicability of LLMs within the finance domain, and we expect continued development and research in this exciting FinTech space. Greater understanding and confidence in model performance and reasoning will increase adoption, and we encourage further transparency from developers regarding model details and the moderation process, including developer biases. LLMs provide exciting new opportunities to extend our understanding of financial and economic mechanisms through enhanced textual analysis, with potential for improved return prediction through better quantification of the information content of text.

[28] For example, a LLM used to assist in financial due-diligence could undermine the process through intrinsic political bias.

References

Aiyappa, R., An, J., Kwak, H., and Ahn, Y.-Y. (2023). "Can we trust the evaluation on ChatGPT?" *arXiv preprint arXiv:2303.12767*.

Araci, D. (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models". *arXiv preprint arXiv:1908.10063*.

Bai, J. J. et al. (2023). "Executives vs. Chatbots: Unmasking Insights through Human-AI Differences in Earnings Conference Q&A". *Available at SSRN*.

Baker, M. and Wurgler, J. (2006). "Investor Sentiment and the Cross-Section of Stock Returns". *Journal of Finance*, 61(4), pp. 1645–1680.

Baker, M. and Wurgler, J. (2007). "Investor Sentiment in the Stock Market". *Journal of Economic Perspectives*, 21(2), pp. 129–151.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). "Measuring economic policy uncertainty". *Quarterly Journal of Economics*, 131(4), pp. 1593–1636.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". *arXiv preprint arXiv:2010.12421*.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). "Bad is Stronger than Good". *Review of General Psychology*, 5(4), pp. 323–370.

Birru, J. and Young, T. (2022). "Sentiment and uncertainty". *Journal of Financial Economics*, 146(3), pp. 1148–1169.

Campbell, J. Y., Grossman, S. J., and Wang, J. (1993). "Trading Volume and Serial Correlation in Stock Returns". *Quarterly Journal of Economics*, 108(4), pp. 905–939.

Campbell, J. Y. and Thompson, S. B. (2008). "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21(4), pp. 1509–1531.

Chen, J., Tang, G., Zhou, G., and Zhu, W. (2023). "ChatGPT, Stock Market Predictability and Links to the Macroeconomy". *Available at SSRN 4660148*.

Clark, T. E. and West, K. D. (2007). "Approximately normal tests for equal predictive accuracy in nested models". *Journal of Econometrics*, 138(1), pp. 291–311.

Da, Z., Engelberg, J., and Gao, P. (2011). "In Search of Attention". *Journal of Finance*, 66(5), pp. 1461–1499.

Da, Z., Engelberg, J., and Gao, P. (2015). "The Sum of All FEARS Investor Sentiment and Asset Prices". *Review of Financial Studies*, 28(1), pp. 1–32.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). "Noise Trader Risk in Financial Markets". *Journal of Political Economy*, 98(4), pp. 703–738.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *arXiv preprint arXiv:1810.04805*.

Federal Reserve Bank of New York. (2023). *Secured Overnight Financing Rate Data*. Federal Reserve Bank of New York. https://www.newyorkfed.org/markets/reference-rates/sofr. Accessed: 1 August 2023.

Federal Reserve Bank of Philadelphia. (2023). *Aruoba-Diebold-Scotti Business Conditions Index [ADS]*. Federal Reserve Bank of Philadelphia. https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/ads. Accessed: 1 August 2023.

Federal Reserve Bank of St. Louis. (2023a). *10-Year Treasury Constant Maturity Minus Federal Funds Rate [T10YFF]*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/T10YFF. Accessed: 1 August 2023.

Federal Reserve Bank of St. Louis. (2023b). *Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity [BAA10Y]*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/BAA10Y. Accessed: 1 August 2023.

Federal Reserve Bank of St. Louis. (2023c). *NBER based Recession Indicators for the United States from the Period following the Peak through the Trough [USREC]*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/USREC. Accessed: 1 August 2023.

Federal Reserve Bank of St. Louis. (2023d). *TED Spread (DISCONTINUED) [TEDRATE]*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/TEDRATE. Accessed: 1 August 2023.

Fieberg, C., Hornuf, L., and Streich, D. (2023). "Using GPT-4 for Financial Advice". *Available at SSRN 4488891*.

Garcia, D. (2013). "Sentiment during Recessions". *Journal of Finance*, 68(3), pp. 1267–1300.

Garcia, D., Hu, X., and Rohrer, M. (2023). "The colour of finance words". *Journal of Financial Economics*, 147(3), pp. 525–549.

Gudibande, A. et al. (2023). "The False Promise of Imitating Proprietary LLMs". *arXiv preprint arXiv:2305.15717*.

Hansen, A. L. and Kazinnik, S. (2023). "Can ChatGPT Decipher Fedspeak?" *Available at SSRN*.

Heston, S. L. and Sinha, N. R. (2017). "News vs. sentiment: Predicting stock returns from news stories". *Financial Analysts Journal*, 73(3), pp. 67–83.

Hirshleifer, D. (2001). "Investor psychology and asset pricing". *Journal of Finance*, 56(4), pp. 1533–1597.

Hutto, C. and Gilbert, E. (2014). "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(1), pp. 216–225.

Jiang, J., Kelly, B., and Xiu, D. (2022). "Expected Returns and Large Language Models". *Available at SSRN*.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). "Predicting returns with text data". (No. w26186) National Bureau of Economic Research.

Li, H. et al. (2023). "Multi-step Jailbreaking Privacy Attacks on ChatGPT". *arXiv preprint arXiv:2304.05197*.

Lopez-Lira, A. and Tang, Y. (2023). "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models". *arXiv preprint arXiv:2304.07619*.

Loughran, T. and McDonald, B. (2011). “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”. *Journal of Finance*, 66(1), pp. 35–65.

Malo, P. et al. (2014). “Good debt or bad debt: Detecting semantic orientations in economic texts”. *Journal of the Association for Information Science and Technology*, 65(4), pp. 782–796.

McGee, R. W. (2023). “Is Chat Gpt Biased Against Conservatives? An Empirical Study”. *Available at SSRN.*

Motoki, F., Neto, V. P., and Rodrigues, V. (2023). “More human than human: measuring ChatGPT political bias”. *Public Choice*, pp. 1–21.

Nakano, M. and Yamaoka, T. (2023). “Enhancing Sentiment Analysis based Investment by Large Language Models in Japanese Stock Market”. *Available at SSRN 4511658.*

Niszczota, P. and Abbas, S. (2023). “GPT has become financially literate: Insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice”. *Finance Research Letters*, 58, p. 104333.

Obaid, K. and Pukthuanthong, K. (2022). “A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news”. *Journal of Financial Economics*, 144(1), pp. 273–297.

Odean, T. (1998). “Volume, Volatility, Price, and Profit When All Traders Are Above Average”. *Journal of Finance*, 53(6), pp. 1887–1934.

Rapach, D. E., Ringgenberg, M. C., and Zhou, G. (2016). “Short interest and aggregate stock returns”. *Journal of Financial Economics*, 121(1), pp. 46–65.

Refinitiv (2021a). *Real-Time News: Feed and Archive - Machine Readable News: Reuters News and Subsets.* English. Version Version 1.1. Refinitiv.

Refinitiv (2021b). *Refinitiv Eikon (with DataStream). Available at: Subscription Service.* Accessed: 1 August 2023.

Romanko, O., Narayan, A., and Kwon, R. H. (2023). “ChatGPT-based Investment Portfolio Selection”. *arXiv preprint arXiv:2308.06260.*

Rozado, D. (2023). “The Political Biases of ChatGPT”. *Social Sciences*, 12(3), p. 148.

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

Tetlock, P., Saar-Tsechansky, M., and Macskassy, S. (2008). “More than words: Quantifying language to measure firms’ fundamentals”. *Journal of Finance*, 63(3), pp. 1437–1467.

Touvron, H. et al. (2023). “LLaMA: Open and Efficient Foundation Language Models”. *arXiv preprint arXiv:2302.13971*.

Welch, I. and Goyal, A. (2008). “A Comprehensive Look at The Empirical Performance of Equity Premium Prediction”. *Review of Financial Studies*, 21(4), pp. 1455–1508.

Wolf, T. et al. (2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45.

Wu, S. et al. (2023). “BloombergGPT: A Large Language Model for Finance”. *arXiv preprint arXiv:2303.17564*.

Yang, Y., UY, M. C. S., and Huang, A. (2020). “FinBERT: A Pretrained Language Model for Financial Communications”. *arXiv preprint arXiv:2006.08097*.

Zaremba, A. and Demir, E. (2023). “ChatGPT: Unlocking the future of NLP in finance”. *Available at SSRN 4323643*.

Zhang, S. et al. (2022). “OPT: Open Pre-trained Transformer Language Models”. *arXiv preprint arXiv:2205.01068*.

Zuckerman, G. (2023). *AI Can Write a Song, but It Can’t Beat the Market, The Wall Street Journal*. https://www.wsj.com/articles/ai-can-write-a-song-but-it-cant-beat-the-market-6df50efd. Accessed: 20 July 2023.

A Appendix

A.1 LLM Recalled Text

To investigate the ability of LLMs ChatGPT and BARD to recall historical daily financial information, we first prompt both LLMs to summarise the most important financial news information for each trading day over the last 20 years.[29] For each trading date in our sample, we submit the following prompt text:

> *"Act as a Financial Analyst. Consider the date XX MONTH XXXX. Provide several paragraphs detailing the most important daily financial news stories relevant to the S&P 500 based on the Thomson Reuters News Archive dataset."*

By instructing the LLMs to act as financial analysts we aim to align the text recalled with the most important information relevant to financial markets. We specify the date format to avoid potential day-month ambiguity, and prompt both LLMs to focus on the Thomson Reuters News Archive as the basis for the response text, as Reuters is a well-known provider of global financial news.[30] Once the responses are collated into two text universes, one for each LLM, we follow the same procedure outlined in Sections 2 and 3, to first derive measures of text sentiment, and then evaluate in-sample and out-of-sample performance.

A.2 LLM Text In-sample Results

In-sample results for daily, weekly, and monthly frequencies over the full dataset are presented in the following section. Results for the ChatGPT and BARD text universes are presented in Panels A and B respectively. Each include a control regression consisting of lagged controls only. Several figures showing daily sentiment significance over time are also presented, for both expanding and rolling 5-year windows of historical observations. All results use robust standard errors.

[29] Each prompt is submitted individually, and in a new session, without prior conversation or priming. To interact with BARD we use the unofficial BARD API (Park, M. [2023]. *BARD API*. https://github.com/dsdanielpark/Bard-API. Accessed: 1 August 2023). Unlike ChatGPT, there were no options to set hyperparameters or fine-tune BARD. Also, BARD returns a default answer, along with two additional answers. We use the default answer. We focus on the U.S. S&P 500 Index since, at the time of writing, both LLMs rate limit answers, which made daily text extraction of all index constituents infeasible. The study period spans over 20 years, from 1st January 2000 to the knowledge cutoff date 30th September 2021 for ChatGPT, and 31st July 2023 for BARD.

[30] The Thomson Reuters News Archives (Refinitiv, 2021a) is an extensive dataset that covers major financial and non-financial news stories globally. Both ChatGPT and BARD have been exposed to the Reuters News Archives, and can recall past articles from this dataset. We focus on this news source in order to reduce potential variability that might occur if no source, or multiple sources were specified. ChatGPT and BARD have been exposed to information from many other leading financial news sources, including the Wall Street Journal, Bloomberg, and the Financial Times to name a few. We focus solely on trading days and exclude weekends and holidays, as these are typically slower news days and hence LLM data coverage of events on these days will likely differ and may complicate text recall.

A.2.1 Returns Forecast - Daily

Panel A: ChatGPT Text Universe

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.0175 *	-0.0059 *	-0.0002 ***	-0.0002 ***	-0.0016 ***	-0.0021 ***
R_t	-0.1079 ***	-0.1088 ***	-0.1087 ***	-0.1091 ***	-0.1095 ***	-0.1099 ***	-0.1093 ***
VIX_t	0.0082	0.0095	0.0093	0.0102	0.0106	0.0116 *	0.0114 *
$Tedrate_t$	-0.0014	-0.0013	-0.0013	-0.0013	-0.0012	-0.0012	-0.0012
$Termspread_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
$BAA10Y_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
Intercept	0.0002	0	-0.0005	-0.0001	-0.0007	-0.0012	-0.0013
R-squared	0.0153	0.0162	0.0162	0.0171	0.0171	0.0179	0.0176
Adj. R-squared	0.0144	0.0151	0.0151	0.016	0.016	0.0168	0.0165
AIC	-32,573	-32,576	-32,576	-32,581	-32,581	-32,585	-32,584
BIC	-32,527	-32,523	-32,523	-32,528	-32,528	-32,532	-32,531
n				5,470			

Panel B: BARD Text Universe

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.0211 ***	-0.0072 **	-0.0002 ***	-0.0002 ***	-0.0012 ***	-0.0019 ***
R_t	-0.0998 ***	-0.1017 ***	-0.101 ***	-0.1014 ***	-0.1019 ***	-0.1029 ***	-0.1021 ***
VIX_t	0.0071	0.0076	0.0082	0.0076	0.0082	0.0092	0.0091
$Tedrate_t$	-0.0015	-0.0011	-0.0013	-0.0013	-0.0012	-0.0009	-0.0009
$Termspread_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0003	-0.0002	-0.0003
$BAA10Y_t$	0	-0.0001	-0.0001	-0.0001	-0.0001	-0.0002	-0.0002
Intercept	-0.0002	0.0002	-0.0005	0.0003	-0.0002	-0.0003	-0.0007
R-squared	0.0133	0.0147	0.0145	0.0144	0.0153	0.0164	0.016
Adj. R-squared	0.0125	0.0137	0.0135	0.0134	0.0143	0.0154	0.015
AIC	-35,273	-35,279	-35,278	-35,278	-35,283	-35,289	-35,287
BIC	-35,226	-35,225	-35,225	-35,224	-35,230	-35,236	-35,233
n				5,930			

Table 9: In-sample daily sentiments and next day market returns - LLM text.
This table reports regression constants β from model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}$, where R_{t+1} is the next day % change of the S&P 500, and $Sent_{t,C}$ is the overall daily pessimism sentiment measure for classifier C, one of the six sentiment classifiers: Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP) and BARD (BA). R_t is the daily % change of the S&P 500, VIX_t is the daily CBOE Volatility Index/100, and $Tedrate$, $Termspread$ and $BAA10Y$ are measures of daily short, medium, and long-term credit risks, as detailed in Section 3.1. The sample period starts 1^{st} January 2000 and ends 30^{th} September 2021 for ChatGPT at knowledge cutoff date, and 31^{st} July 2023 for BARD. Robust standard errors are used. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 9 reports in-sample daily results. Consistent with prior literature, and news sentiment results Table 3, next day return reversals are observed for all classifiers for both text universes. All sentiments are negative and significant, with improved adj. R^2 values greater than the control model. ChatGPT outperforms BARD in both Panels A and B, with the highest adj. R^2 and BIC values, followed by Bard, then transformers and dictionaries last. Both ChatGPT and BARD have higher BIC values than the control model, while most other models are slightly less, implying the inclusion of ChatGPT and BARD sentiments contributes a marginal amount of new information.[31]

[31] While the BARD sample is roughly 10% larger, using the same end-date 30^{th} September 2021 makes no difference to the main results, the only notable difference being adj. R^2 values are closer.

A.2.2 Returns Forecast - Weekly

Panel A: ChatGPT Text Universe

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.1566 **	-0.0623 ***	-0.0013 ***	-0.0017 ***	-0.0105 ***	-0.0144 ***
R_t	-0.0773	-0.0824 *	-0.0847 *	-0.0837 *	-0.0854 *	-0.087 *	-0.0861 *
VIX_t	0.0246	0.0352	0.035	0.0363	0.0411 *	0.0445 *	0.045 *
$Tedrate_t$	-0.009 *	-0.0075	-0.008	-0.0083	-0.0077	-0.0075	-0.0074
$Termspread_t$	-0.0014	-0.0014	-0.0013	-0.0016	-0.0017 *	-0.0016	-0.0017 *
$BAA10Y_t$	0.0009	0.0014	0.0018	0.0014	0.0012	0.0013	0.0013
Intercept	0.0002	-0.0016	-0.0074	-0.0017	-0.0067	-0.0085	-0.0101 *
R-squared	0.0218	0.0308	0.0324	0.0315	0.0352	0.037	0.038
Adj. R-squared	0.0175	0.0256	0.0272	0.0264	0.0301	0.0318	0.0329
AIC	-5,166	-5,174	-5,176	-5,175	-5,179	-5,181	-5,183
BIC	-5,131	-5,134	-5,136	-5,135	-5,139	-5,141	-5,142
n				1,135			

Panel B: BARD Text Universe

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.1565 ***	-0.0605 ***	-0.0017 ***	-0.0018 ***	-0.0082 ***	-0.0106 ***
R_t	-0.082 *	-0.0903 **	-0.0879 *	-0.094 **	-0.0941 **	-0.0941 **	-0.0908 **
VIX_t	0.017	0.0202	0.0253	0.0209	0.0239	0.0294	0.0275
$Tedrate_t$	-0.0088 *	-0.0059	-0.0069	-0.0067	-0.0064	-0.0051	-0.0054
$Termspread_t$	-0.0013 *	-0.0014 *	-0.0012	-0.002 ***	-0.002 ***	-0.0018 **	-0.0018 **
$BAA10Y_t$	0.0018	0.0016	0.0017	0.0016	0.0014	0.0007	0.001
Intercept	-0.0013	0.0016	-0.0045	0.003	-0.0019	-0.0023	-0.0043
R-squared	0.0215	0.0295	0.0306	0.0334	0.0348	0.0391	0.0328
Adj. R-squared	0.0175	0.0248	0.0259	0.0286	0.03	0.0344	0.0281
AIC	-5,593	-5,601	-5,603	-5,606	-5,608	-5,613	-5,605
BIC	-5,557	-5,560	-5,562	-5,565	-5,567	-5,572	-5,564
n				1,231			

Table 10: In-sample weekly sentiments and next week market returns - LLM text.
This table reports regression constants β from model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}$, where R_{t+1} is the % change of the S&P 500 for the following week, $Sent_{t,C}$ is the average daily pessimism sentiment measure for classifier C for the prior week, R_t is the % change of the S&P 500 for the prior week, VIX_t is the CBOE Volatility Index/100 as of market close of the prior week, and $Tedrate$, $Termspread$ and $BAA10Y$ are measures of daily short, medium, and long-term credit risks, taken at market close of the prior week, as detailed in Section 3.1. The sample period starts 1^{st} January 2000 and ends 30^{th} September 2021 for ChatGPT at knowledge cutoff date, and 31^{st} July 2023 for BARD. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Weekly in-sample results Table 10 are similar to those of the daily frequency LLM models, with negative and significant sentiment coefficients. All sentiment models have improved adj. R^2 and BIC values compared to controls, with both ChatGPT and BARD the best performers. Compared to daily LLM results, the inclusion of weekly lagged LLM sentiments lead to much larger relative increases in adj. R^2 values, suggesting the recalled text of both LLMs is capturing information at the weekly frequency in addition to the daily frequency. For completeness, a weekly regression model including all sentiments simultaneously is presented in Table 14, with only GP sentiment significant for the BARD text universe.

A.2.3 Returns Forecast - Monthly

Panel A: ChatGPT Text Universe

Variables	Control_1	Control_2	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-	-0.241	-0.1249	-0.0014	-0.0018	-0.0098	-0.0074
BW_t	-	-0.0041	-0.003	-0.0025	-0.004	-0.0037	-0.0039	-0.0043
dp_t	0.0182	0.021	0.0306	0.0422	0.0291	0.0335	0.034	0.0246
dy_t	0.1007	0.0884	0.0695	0.0643	0.0721	0.0673	0.0677	0.0769
ep_t	-0.0028	-0.0029	-0.0003	0.0015	-0.0018	-0.0017	-0.0017	-0.0025
$svar_t$	0.2157	0.1947	0.2532	0.228	0.2279	0.2601	0.2686	0.242
bm_t	-0.0322	-0.0398	-0.036	-0.051	-0.0425	-0.044	-0.044	-0.039
$ntis_t$	0.7776 **	0.7672 **	0.7242 **	0.748 **	0.7408 **	0.7363 **	0.7391 **	0.7408 **
ltr_t	0.0167	0.0094	0.0273	0.0289	0.0169	0.024	0.0231	0.0166
tms_t	-0.5471 **	-0.5643 **	-0.5009 *	-0.4428	-0.5203 *	-0.5217 *	-0.5228 *	-0.5429 *
dfy_t	-2.0533	-1.9775	-1.3755	-1.234	-1.6345	-1.5182	-1.5176	-1.6929
dfr_t	1.5622 **	1.7929 **	1.7724 **	1.9819 **	1.8825 **	1.7956 **	1.795 **	1.7775 **
$infl_t$	1.0243	1.0459	1.0479	1.0719	1.0077	1.0419	1.0571	1.0231
Intercept	0.4982 **	0.4595 **	0.4248 *	0.4467 *	0.4273 *	0.4213 *	0.4243 *	0.4231 *
R-squared	0.1142	0.1156	0.1167	0.1212	0.1145	0.1156	0.1152	0.1131
Adj. R-squared	0.0751	0.0728	0.07	0.0748	0.0677	0.0689	0.0685	0.0663
AIC	-904.7	-903.1	-898.5	-899.9	-897.9	-898.2	-898.1	-897.5
BIC	-858.3	-853.2	-845.1	-846.5	-844.5	-844.8	-844.7	-844.1
n	260							

Panel B: BARD Text Universe

Variables	Control_1	Control_2	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-	-0.6178 ***	-0.2419 ***	-0.0058 ***	-0.0055 **	-0.0234 ***	-0.0371 ***
BW_t	-	-0.0057	-0.0026	-0.0033	-0.0032	-0.004	-0.004	-0.0028
dp_t	0.0546	0.058	0.0767	0.0872	0.0974	0.091	0.0917	0.0854
dy_t	0.068	0.0501	0.011	0.0156	-0.002	-0.002	-0.012	-0.0007
ep_t	-0.0059	-0.0059	-0.0007	0.0017	-0.0021	-0.001	-0.001	-0.0008
$svar_t$	0.0422	0.0254	0.2211	0.2004	0.1476	0.162	0.3102	0.3851
bm_t	-0.0335	-0.0448	-0.0445	-0.0913	-0.0549	-0.0655	-0.0676	-0.0709
$ntis_t$	0.8212 **	0.8111 **	0.6356 *	0.7517 **	0.7409 **	0.7387 **	0.6816 **	0.6569 *
ltr_t	0.0189	0.0051	0.0769	0.0612	0.0381	0.0487	0.0659	0.0688
tms_t	-0.5957 **	-0.6209 **	-0.5722 **	-0.4679 *	-0.752 ***	-0.7048 **	-0.669 **	-0.6405 **
dfy_t	-2.1956	-2.0694	-0.8248	-0.8178	-1.2422	-0.9623	-0.7018	-0.7013
dfr_t	1.3307 *	1.6729 **	0.822	1.4278 *	1.2146	1.1608	0.826	0.7351
$infl_t$	0.4882	0.5596	0.7192	0.8188	0.5348	0.6097	0.6373	0.6861
Intercept	0.5105 **	0.4517 *	0.3965 *	0.4488 *	0.4287 *	0.3904	0.3526	0.3677
R-squared	0.1089	0.1118	0.138	0.1409	0.1364	0.1357	0.141	0.1418
Adj. R-squared	0.0706	0.07	0.0939	0.097	0.0922	0.0915	0.0971	0.0979
AIC	-922.1	-921.0	-927.0	-927.9	-926.5	-926.3	-927.9	-928.2
BIC	-875.4	-870.7	-873.1	-874.0	-872.6	-872.4	-874.1	-874.3
n	268							

Table 11: In-sample monthly sentiments and next month market returns - LLM text.
This table reports regression constants β for monthly regression model Equation (2), which fits next month S&P 500 returns, $CRSP\_SPvw_{t+1}$, from prior month text sentiment and controls. We include controls from Welch and Goyal (2008), which are better suited to capturing monthly frequency effects than daily and weekly model controls. Two measures of sentiment are included; $Sent_t$ is the average monthly sentiment derived from LLM text, and BW_t is the well-known Baker and Wurgler (2007) sentiment index. Control variable details are provided in Appendix A.6. Two control models are included for clarity. The first is comprised of all controls excluding both text-based sentiment and the Baker Wurgler sentiment index, while the second excludes only text-based sentiment. The sample period starts 1st January 2000 for both text universes, and ends 30th September 2021 for ChatGPT at knowledge cutoff date, and 30th June 2022 for BARD, since the Baker Wurgler sentiment index is currently available up to this date. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

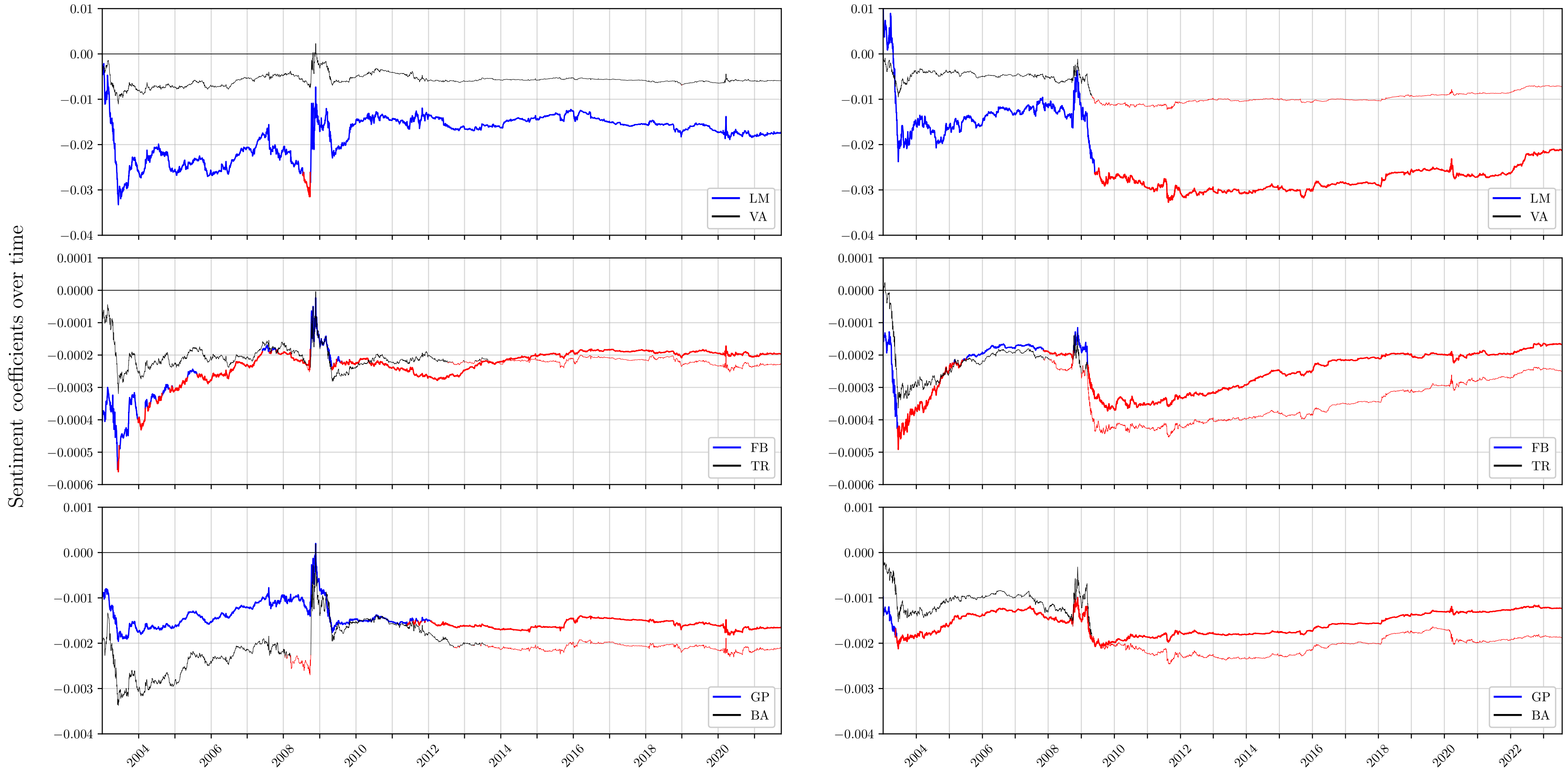
Monthly in-sample results are reported in Table 11. There is a notable difference between ChatGPT and BARD text at this frequency, as although all text-derived sentiments are negative, they are only significant for the BARD text universe, where they are much larger. Overall, results suggest the inclusion of sentiment does not improve model performance, as the inclusion of sentiment only results in gains to adj. R^2 values for BARD text, but all BIC magnitudes across both panels are smaller than the control models. Further, Baker Wurgler sentiment is negative, but not significant, and there is no improvement between control models 1 and 2. These results support prior literature that Baker Wurgler sentiment is a contemporaneous explanatory variable rather than a predictor of future returns, (Baker and Wurgler, 2007).

It is worth noting due to the lower monthly frequency, there are less observations than for both daily and weekly models. Combined with the increased number of independent variables, the reported significance of text-based sentiments for BARD text should be viewed cautiously, as reduced sample size in combination with the greater number of control variables increases the chance of overfitting. For completeness, we present monthly models with sentiments from all classifiers in Appendix A.4.

A.2.4 LLM Sentiment Significance Over Time

Results for sentiment significance over time using both expanding window and 5-year rolling window approaches for ChatGPT and BARD text universes are presented in Figures 5 and 6. Note the last value of the expanding window series corresponds to the full in-sample daily coefficients of Table 9.

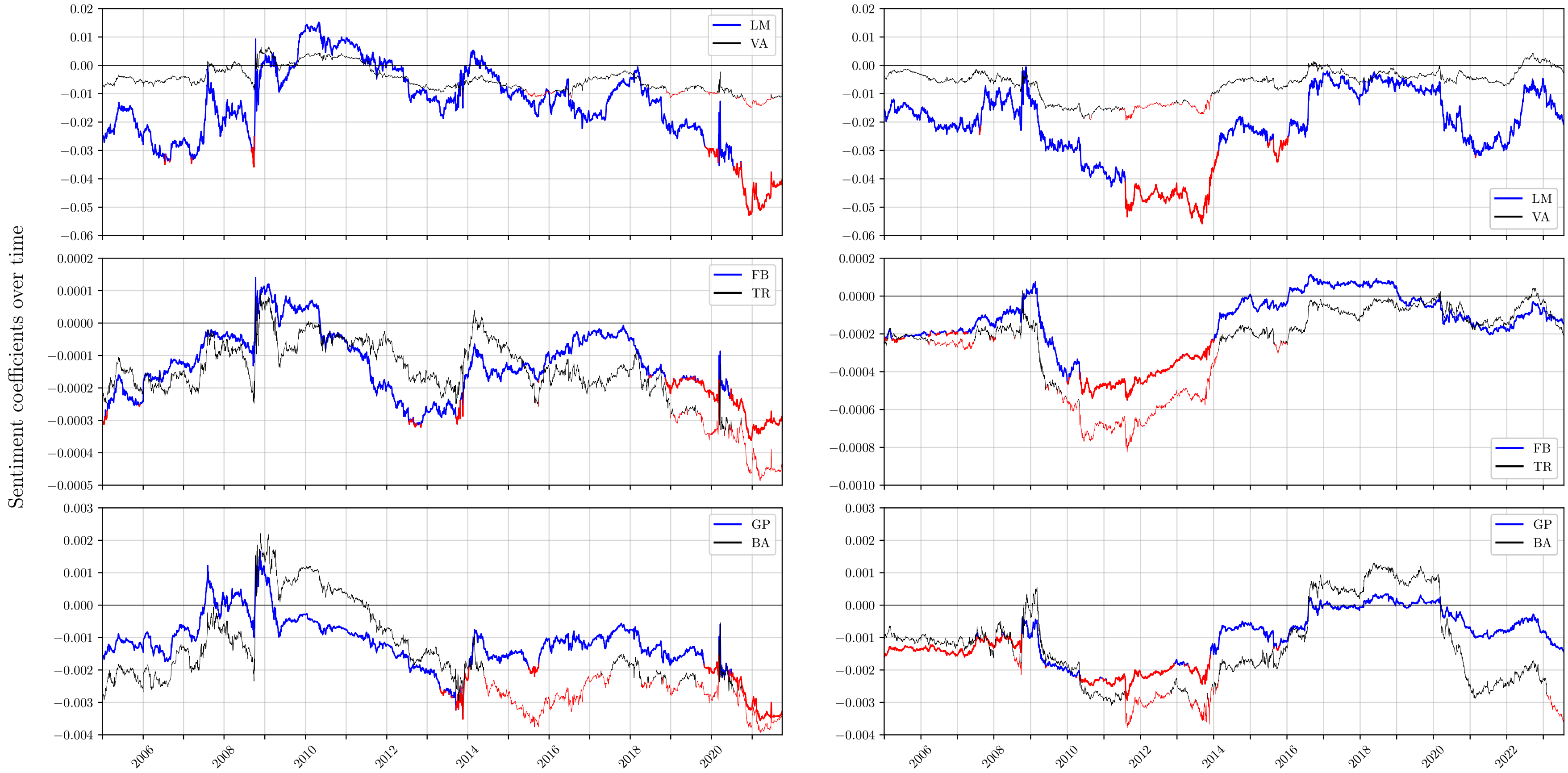
Figure 5 shows sentiment coefficients remain consistently negative over time when an expanding window of observations is used. While sentiment coefficients are similar across both text universes, sentiments from ChatGPT and BARD are significant for longer periods of time than both simpler transformers and dictionary methods. In addition, sentiments derived from BARD text are significant for longer periods of time relative to those derived from ChatGPT text. For example, ChatGPT sentiment from BARD text is significant for almost the full sample at the 5% level, even during the GFC of 2008-2009. These results suggest that, assuming behavioural models of next-day return reversals hold, both ChatGPT and BARD outperform both simpler transformers and dictionary methods in sentiment classification. Values for all models mostly stabilize after roughly 2004, suggesting our choice of 3 years for initial coefficient estimates is appropriate.



Panel A: ChatGPT Text Universe

Panel B: BARD Text Universe

Figure 5: Sentiment Coefficients over Expanding Window - ChatGPT and BARD Text.
These graphs show sentiment coefficient values β_1 of sentiment models Equation (1), calculated over an expanding window of historical observations from 1st January 2000 up to time t. Panel A and B report results for ChatGPT and BARD text universes respectively. The first three years of the sample are used to calculate initial values. The six sentiment measures are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP) and BARD (BA). Classifiers are grouped in pairs; dictionaries, transformers, and LLMs. The sample period ends 30th September 2021 for Panel A, and 31st July 2023 for Panel B. Values are highlighted in red if significant at time t, for $p < 0.05$ based on robust standard errors.



Panel A: ChatGPT Text Universe

Panel B: BARD Text Universe

Figure 6: Sentiment Coefficients over 5-Year Rolling Window - ChatGPT and BARD Text.
These graphs show sentiment coefficient values β_1 of sentiment models Equation (1), calculated over a 5-year rolling window of historical sentiment, controls, and returns. The first 5 years of the sample are used to calculate initial values. The six sentiment measures are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP), and BARD (BA). Classifiers are grouped in pairs; dictionaries, transformers, and LLMs. The sample period ends 30th September 2021 for Panel A, and 31st July 2023 for Panel B. Values are highlighted in red if significant at time t, for $p < 0.05$ based on robust standard errors.

A.3 LLM Text Out-of-sample Results

A.3.1 Out-of-sample R^2, CER, and Sharpe Ratios

Out-of-sample results for both LLM text universes are detailed in Table 12, following the same methodology detailed in Section 3.2. Results show ChatGPT and BARD sentiment models outperform all others across both text universes, performing the best and second best respectively, with the largest R^2_{OOS} values, while dictionary models LM and VA have the smallest.

Panel A: ChatGPT Text Universe 1-1-2000 to 30-9-2021

	R^2_{OOS} (%)		CER gain (%)			CER (%)		Sharpe ratios		
Classifier	$Sent._{Cont.}$	$Sent._{Hist.}$	$Sent._{Hist.}$	$Sent._{Cont.}$	$Cont._{Hist.}$	Sent.	Cont.	Sent.	Cont.	Hist.
LM	0.02	0.67***	5.787	0.953	4.834	8.194	7.241	0.761	0.696	0.398
VA	0.00	0.65***	6.244	1.409		8.650		0.802		
FB	0.08***	0.73***	5.850	1.015		8.256		0.763		
TR	0.11**	0.77***	6.075	1.241		8.482		0.784		
GP	**0.20*****	**0.85*****	**6.875**	**2.041**		**9.282**		**0.851**		
BA	0.14**	0.80***	6.803	1.969		9.210		0.830		

Panel B: BARD Text Universe 1-1-2000 to 31-7-2023

	R^2_{OOS} (%)		CER gain (%)			CER (%)		Sharpe ratios		
Classifier	$Sent._{Cont.}$	$Sent._{Hist.}$	$Sent._{Hist.}$	$Sent._{Cont.}$	$Cont._{Hist.}$	Sent.	Cont.	Sent.	Cont.	Hist.
LM	0.10***	0.58***	6.249	1.266	4.983	8.545	7.280	0.769	0.683	0.330
VA	0.08**	0.56***	6.155	1.172		8.452		0.764		
FB	0.08***	0.56***	6.301	1.318		8.597		0.768		
TR	0.18***	0.66***	7.201	2.218		9.497		0.829		
GP	**0.29*****	**0.77*****	**8.707**	**3.724**		**11.004**		**0.931**		
BA	0.25***	0.73***	7.833	2.849		10.129		0.878		

Table 12: Out-of-sample analysis results - LLM text.
Summary out-of-sample results for daily sentiment models which forecast next day % returns of the S&P 500 index from an expanding window of observations, according to Equation (1). The six sentiment models are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP) and BARD (BA). Two sets of R^2_{oos} are presented; sentiment model forecasts vs control model forecasts, and sentiment model forecasts vs historical mean returns model forecasts. Significance of the right-tailed nested model test of Clark and West (2007) is shown for $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.[32] Certainty equivalent return (CER) % and CER gain % values of portfolios based on both sentiment model forecasts and benchmark model forecasts are presented, calculated per Section 3.2 and annualized.[33] Three sets of CER gain values are presented for clarity; CER gain of sentiment portfolios over historical mean returns portfolio, CER gain of sentiment portfolios over the control portfolio, and the CER gain of the control portfolio over the historical mean returns portfolio. Raw CER values for both sentiment and control portfolios are shown. Annualized Sharpe ratios for sentiment portfolios, and both benchmark portfolios are presented.[34] The sample period starts 1st January 2000 for both text universes, and ends 30th September 2021 for ChatGPT at knowledge cutoff date, and 31st July 2023 for BARD. R^2_{oos} is calculated from all data 1st January 2003 onwards, while CER, CER gain, and Sharpe ratios are calculated from all data 1st January 2006 onwards.

One interesting aspect of the results is that R^2_{OOS} values for sentiment models vs control model

[32] Clark and West equal predictive performance test of nested models. We use the right-tailed test, which tests the null hypothesis of equal model performance against the alternative that the larger sentiment model outperforms the nested model in terms of MSPE.

[33] CER and CER gain values are annualized by multiplying by 250.

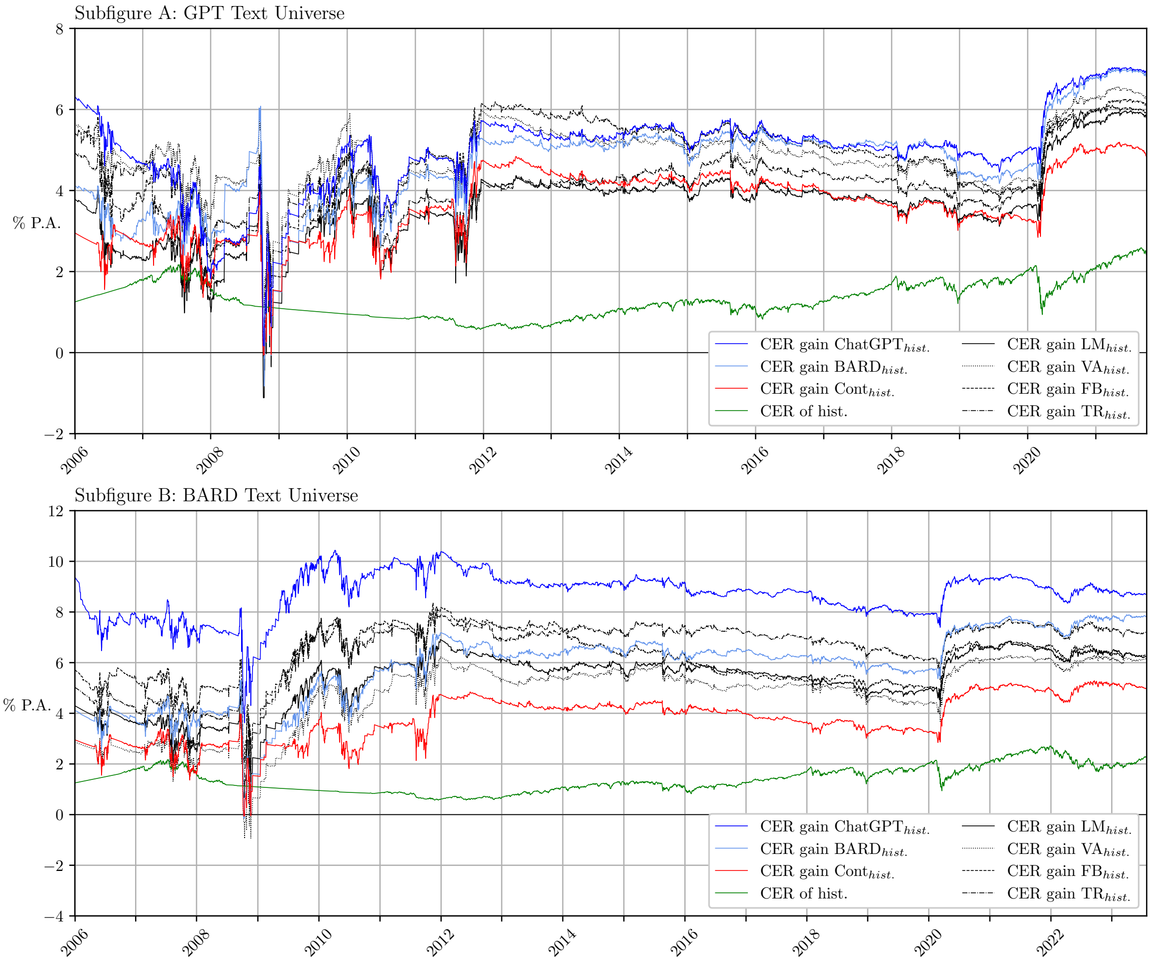
[34] Sharpe ratios are calculated from daily portfolio performance per Section 3.2. The daily risk-free rate is calculated as the daily compounding equivalent of the 3-Month Treasury Bill rate [DTB3], see (Federal Reserve Bank of St. Louis., 2023d). Since the resulting Sharpe ratios are daily, we annualize by multiplying by $\sqrt{250}$.

are higher for BARD text than ChatGPT text, however this pattern reverses for sentiment models vs historical mean model. The difference between both benchmarks suggests control variables account for an R^2_{OOS} improvement above the historical mean of roughly 0.65% for ChatGPT text, but only 0.48% for BARD text. At the same time, CER and CER gain values of the control portfolio suggests marginally better performance vs the historical mean returns for BARD text, in apparent conflict with R^2_{OOS} results. Sharpe ratios offer further insight, with BARD text having greater sentiment portfolio performance than ChatGPT text, but worse control and means portfolio performance. Crucially, the difference in historical means performance between ChatGPT and BARD text is larger (0.068) than the difference in controls (0.013). Hence part of the larger CER gains observed for BARD text come from the relatively larger drop in performance of the historical means portfolio, as this contributes to inflated CER gain values. The most likely reason why BARD text observes lower R^2_{OOS} values than ChatGPT text for sentiment models compared against historical mean returns model is because of changes in sentiment performance particularly late in the sample, possibly post-ChatGPT knowledge cutoff.[35] Still, the fact sentiment models for the BARD text universe consistently result in outperformance relative to the ChatGPT text universe is further evidence BARD text may be capturing some unique information not captured by ChatGPT text. Further, ChatGPT forecast models and portfolios consistently outperform all other models and portfolios across both Panels A and B, suggesting ChatGPT is better able to identify and classify this information than BARD.

A.3.2 CER over time

CER gain values over time for both text universes are presented in Figure 7. Results show sentiment-based portfolios consistently outperform both the control portfolio (red), and the historical mean returns benchmark (green) over the full sample. ChatGPT outperforms all other classifiers across both text universes, and the CER gains of all sentiment portfolios exceed those of the control portfolio, demonstrating the inclusion of sentiment leads to better portfolio performance.

[35] We would require several additional years of observations to determine if this is the case. For comparison the post-ChatGPT knowledge cutoff period to end of BARD sample is just over half of the 3-year initial estimation period used for both initial model estimates and CER calculations. Hence this period is too short to draw any strong conclusions.


Subfigure A: GPT Text Universe

Subfigure B: BARD Text Universe

Figure 7: CER gain values over time - ChatGPT and BARD Text Universes.
These graphs show annualized changes in CER gain values of daily sentiment portfolios vs the historical mean returns portfolio over time. The two best performers are ChatGPT (blue) and BARD (light blue). Daily portfolio allocations are made between the S&P 500 Index and the daily risk-free rate based on next day model forecasts over an expanding window of observations. An initial 3-year estimation period is used for initial model forecasts and initial next day portfolio allocations, (4). CER gain values are then calculated based on an expanding window of daily portfolio returns, with 3 years of initial trading used to calculate initial CER values, (5). Hence CER gain values are presented from 1^{st} January 2006 onwards, with values at time t representing CER gain of portfolios trading from 1^{st} January 2003 to t. For completeness, we include the CER gain of the control portfolio above the historical mean returns model (red), and the CER of the historical mean returns model (green), both of which are identical for both text universes. Note the last value of each series correspond to full sample results of Table 12, with sample period ending 30^{th} September 2021 and 31^{st} July 2023 for ChatGPT and BARD text universes respectively.

A.4 Returns forecasts – All Sentiments: Daily, Weekly, Monthly

To check if any one classifier dominates the others head-to-head, we conduct in-sample OLS regressions that include all sentiment measures simultaneously, along with controls. Daily, weekly, and monthly results for both news summary texts, as well as both LLM text universes, are presented in Tables 13, 14, and 15 respectively.

	Daily Frequency					
	Panel A: News Text		Panel B: ChatGPT Text		Panel C: BARD Text	
Variables	Control	Full Model	Control	Full Model	Control	Full Model
LM_t	-	0.0024	-	0.013	-	0.0001
VA_t	-	-0.004	-	0.0005	-	0.0001
FB_t	-	0.0002	-	-0.0001	-	0.0001
TR_t	-	-0.0002	-	0	-	0
GP_t	-	-0.0014 *	-	-0.0013	-	-0.001
BA_t	-	-	-	-0.0009	-	-0.0008
R_t	-0.1155 ***	-0.1622 ***	-0.1079 ***	-0.1098 ***	-0.0998 ***	-0.1027 ***
VIX_t	0.0067	0.008	0.0082	0.012 *	0.0071	0.0094 *
$TedRate_t$	-0.0014	-0.0015	-0.0014	-0.0012	-0.0015	-0.0008
$TermSpread_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
$BAA10Y_t$	-0.0002	-0.0001	-0.0002	-0.0002	0	-0.0003
Intercept	0.0002	-0.0008	0.0002	-0.0014	-0.0002	-0.0006
R-squared	0.0162	0.0199	0.0153	0.0182	0.0133	0.0165
Adj. R-squared	0.0152	0.0179	0.0144	0.0163	0.0125	0.0147
AIC	-29,420	-29,429	-32,573	-32,577	-35,273	-35,280
BIC	-29,374	-29,351	-32,527	-32,491	-35,226	-35,193
n	4,958		5,417		5,930	

Table 13: In-sample daily full sentiment models.
This table reports regression constants β from model: $R_{t+1} = \alpha + \sum_{i=1}^{6} \beta_i Sent_{t,C} + \beta_7 R_t + \beta_8 VIX_t + \beta_9 Tedrate_t + \beta_{10} Termspread_t + \beta_{11} BAA10Y_t + \varepsilon_{t+1}$, at the daily frequency, where each classifiers sentiment measure is included, i.e. $\sum_{i=1}^{6} \beta_i Sent_{t,C}$. Results for news text, ChatGPT recalled text, and BARD recalled text are presented in Panels A, B, C respectively. All sample periods start 1^{st} January 2000, and end 31^{st} July 2020 for Panel A, 30^{th} September 2021 for Panel B at ChatGPT knowledge cutoff date, and 31^{st} July 2023 for Panel C. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	Weekly Frequency					
	Panel A: News Text		Panel B: ChatGPT Text		Panel C: BARD Text	
Variables	Control	Full Model	Control	Full Model	Control	Full Model
LM_t	-	-0.2159 **	-	0.1477	-	0.095
VA_t	-	0.015	-	-0.0285	-	-0.0178
FB_t	-	-0.0001	-	0.0007	-	-0.0004
TR_t	-	-0.0012	-	-0.0006	-	0.0000
GP_t	-	0.0116 *	-	-0.005	-	-0.0158 ***
BA_t	-	-	-	-0.0135	-	0.0129
R_t	-0.0768	-0.0694	-0.0773	-0.0881 *	-0.082 *	-0.094 **
VIX_t	0.0208	0.0307	0.0246	0.0476 **	0.017	0.0294
$TedRate_t$	-0.0088	-0.01 *	-0.009 *	-0.0077	-0.0088 *	-0.0065
$TermSpread_t$	-0.0014	-0.0021 *	-0.0014	-0.0017 *	-0.0013 *	-0.0019 **
$BAA10Y_t$	0.0012	0.0011	0.0009	0.0013	0.0018	0.0007
Intercept	0.0000	0.0034	0.0002	-0.0169 **	-0.0013	-0.0012
R-squared	0.0199	0.0303	0.0218	0.0401	0.0215	0.0422
Adj. R-squared	0.0153	0.0211	0.0175	0.0307	0.0175	0.0335
AIC	-4,866	-4,850	-5,166	-5,175	-5,593	-5,607
BIC	-4,831	-4,791	-5,131	-5,110	-5,557	-5,541
n	1,072		1,135		1,231	

Table 14: In-sample weekly full sentiment models.
This table reports regression constants β from model: $R_{t+1} = \alpha + \sum_{i=1}^{6} \beta_i Sent_{t,C} + \beta_7 R_t + \beta_8 VIX_t + \beta_9 Tedrate_t + \beta_{10} Termspread_t + \beta_{11} BAA10Y_t + \varepsilon_{t+1}$, at the weekly frequency, where each classifiers sentiment measure is included, i.e. $\sum_{i=1}^{6} \beta_i Sent_{t,C}$. Results for news text, ChatGPT recalled text, and BARD recalled text are presented in Panels A, B, C respectively. All sample periods start 1^{st} January 2000, and end 31^{st} July 2020 for Panel A, 30^{th} September 2021 for Panel B at ChatGPT knowledge cutoff date, and 31^{st} July 2023 for Panel C. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

	Monthly Frquency								
	Panel A: News Text			Panel B: ChatGPT Text			Panel C: BARD Text		
Variables	Control_1	Control_2	Full Model	Control_1	Control_2	Full Model	Control_1	Control_2	Full Model
LM_t	-	-	-0.0321	-	-	0.1602	-	-	0.0823
VA_t	-	-	0.5762 ***	-	-	-0.3647 **	-	-	-0.1625
FB_t	-	-	0.0092	-	-	-0.0026	-	-	-0.0046
TR_t	-	-	0	-	-	0.0014	-	-	0.0087
GP_t	-	-	-0.0753 **	-	-	-0.0729	-	-	-0.0138
BA_t	-	-	-	-	-	0.1424 **	-	-	-0.0253
BW_t	-	-0.0044	-0.0078	-	-0.0041	-0.0017	-	-0.0057	-0.0022
dp_t	0.0114	0.0152	0.0015	0.0182	0.021	0.0555	0.0546	0.058	0.0923
dy_t	0.1007	0.0866	0.0951	0.1007	0.0884	0.0928	0.068	0.0501	0.0025
ep_t	-0.0021	-0.0023	-0.0096	-0.0028	-0.0029	0.0044	-0.0059	-0.0059	0.0001
$svar_t$	0.222	0.1952	0.159	0.2157	0.1947	0.0598	0.0422	0.0254	0.41
bm_t	-0.0214	-0.0284	0.0037	-0.0322	-0.0398	-0.1014	-0.0335	-0.0448	-0.0825
$ntis_t$	0.71 **	0.6967 **	0.3558	0.7776 **	0.7672 **	0.924 ***	0.8212 **	0.8111 **	0.6725 *
ltr_t	0.0326	0.0233	0.0366	0.0167	0.0094	0.0011	0.0189	0.0051	0.0694
tms_t	-0.4896 *	-0.5121 *	-0.549 *	-0.5471 **	-0.5643 **	0	-0.5957 **	-0.6209 **	-0.5362 *
dfy_t	-2.0185	-1.9422	-2.6324 *	-2	-2	-2	-2.1956	-2.0694	-0.7581
dfr_t	1.3593 *	1.613 **	0.6036	1.5622 **	1.7929 **	2.6162 ***	1.3307 *	1.6729 **	0.9335
$infl_t$	0.889	0.8961	0.3993	1.0243	1.0459	1.1294	0.4882	0.5596	0.7461
Intercept	0.4701 **	0.4272 *	0.3838 *	0.4982 **	0.4595 **	0.6538 **	0.5105 **	0.4517 *	0.4202
R-squared	0.1104	0.112	0.1563	0.1142	0.1156	0.1411	0.1089	0.1118	0.147
Adj. R-squared	0.0687	0.0665	0.0934	0.0751	0.0728	0.0769	0.0706	0.07	0.0854
AIC	-854.4	-852.9	-852.5	-904.7	-903.1	-895.8	-922.1	-921.0	-919.8
BIC	-808.8	-803.7	-785.9	-858.3	-853.2	-824.6	-875.4	-870.7	-848.0
n		247			260			268	

Table 15: In-sample monthly full sentiment models.
This table reports monthly regression constants β from Equation (2), where $Sent_{t,C}$ is replaced with sentiment from each classifier, i.e. $\sum_{i=1}^{6} \beta_i Sent_{t,C}$. See Appendix A.6 for control variable details. Results for news text, ChatGPT recalled text, and BARD recalled text are presented in Panels A, B, C respectively. All sample periods start 1^{st} January 2000, and end 31^{st} July 2020 for Panel A, 30^{th} September 2021 for Panel B at ChatGPT knowledge cutoff date, and 30^{th} June 2022 for Panel C at the last available Baker Wurgler sentiment index value. Robust standard errors are used. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The results of Tables 13, 14, and 15 show no one sentiment classifier consistently dominates the others, likely in part due to collinearity between multiple sentiments reducing overall significance of any one classifier. The majority of sentiment coefficients are negative, and there is some evidence that ChatGPT uniquely identifies information on the lower weekly frequency, as this is present for both news and BARD text universes. However despite improved adj. R^2_{oos} values for almost all models, they are all accompanied by worse BIC values than the control models. Further, the adj. R^2_{oos} are worse than the majority of individual in-sample sentiment models, hence there is little benefit in including multiple sentiments simultaneously.

A.5 LM Negate Keywords

The following negation keywords were used to change an otherwise positive word count to a negative word count for the Loughran & McDonald Dictionary classifier, LM:

aint, arent, cannot, cant, couldnt, darent, didnt, doesnt, ain't, aren't, can't, couldn't, daren't, didn't, doesn't, dont, hadnt, hasnt, havent, isnt, mightnt, mustnt, neither, don't, hadn't, hasn't, haven't, isn't, mightn't, mustn't, neednt, needn't, never, none, nope, nor, not, nothing, nowhere, oughtnt, shant, shouldnt, wasnt, werent, oughtn't, shan't, shouldn't, wasn't, weren't, without, wont, wouldnt, won't, wouldn't, rarely, seldom, despite, no, nobody.

A.6 Monthly Frequency Control Variables

Monthly controls included the Baker and Wurgler (2007) Sentiment Index , along with variables from Welch and Goyal (2008). Brief descriptions are provided below, see the original papers for further details:

BW: Baker Wurgler SENT index.

dp: Dividend Price Ratio, the difference between log of dividends and log of prices.

dy: Dividend Yield, the difference between log of dividends and log of lagged prices.

ep: Earnings Price Ratio, difference between log of earnings and log of prices.

svar: Stock Variance, computed as the sum of squared daily returns of the S&P 500.

bm: Book-to-Market Ratio, the ratio of book value to market value for the Dow Jones Industrial Average.

ntis: Net Equity Expansion, ratio of 12-month moving sums of net issues by NYSE listed stocks divided by total end-of-year market capitalisation of NYSE stocks.

ltr: Long Term Rate of Returns.

tms: Term Spread, the difference between the long term yield on government bonds [lty] and the Treasury-bill [tbl].

dfy: Default Yield Spread, difference between BAA and AAA corporate bond yields.

dfr: Default Return Spread, the difference between long-term corporate bond and long-term government bond returns.

infl: Inflation, the Consumer Price Index.

B Appendix

B.1 Parsing ChatGPT Classification Text

When using ChatGPT to classify ChatGPT text summaries, the responses were almost always a bullet list in the format *class: score*. All but one response had this same text format or an almost identical format, which made parsing simple. The one odd response had three sub-ratings under positivity, one each for; S&P 500 Index, General Electric, and Oracle Corp. For this day the positive sentiment measure was taken as the average of these three values, divided by 100.

When using ChatGPT to classify BARD text summaries, the responses were more varied. Responses generally consisted of bullet formats similar to ChatGPT classification, but several different response formats were also returned. These sometimes included full explanatory paragraphs after, or between the sentiment classes and scores. In addition, for two days scores of 'None' and 'N/A' were returned instead of 1 for the lowest score as prompted. Regex was used to parse these different text patterns, extract sentiment scores, and to replace 'None' and 'N/A' with 1 where applicable. A remaining 62 daily summaries consisted of either days with multiple sub-ratings for either positive or negative, or a response with only explanatory text and no numerical sentiment scores. For the first case averages were used, while for the second case these queries were re-run to obtain numerical sentiment scores. We confirmed there were no sentiment score outliers, then divide each day by 100 and save the results.

B.2 Parsing BARD Classification Text

When using BARD to classify ChatGPT text summaries, the responses consisted of several different patterns. Most consisted of some type of format *class: score*. However, many answers had varying formats, with different spacing, separating characters, additional text before, after, or between the classifications. Hence BARD responses required extensive text parsing to extract the sentiment scores from each response. Regex was used to screen for the eight main patterns present, and to overcome subtle differences in the text patterns. Due to overlapping pattern similarities in the output text, parsing was particularly challenging, as we strike a balance between specifying unique parsing patterns, while not being overly-repetitive. Through careful use of regex and parsing order, we balanced this trade-off, and parse most cases. Eight cases in which BARD misspelled ‘positivity’ were also corrected. Four remaining entries with unique patterns were manually parsed, one outlier corrected, and 185 entries with zero recorded for either positive or negative were corrected to 1 for consistency. We confirmed there were no sentiment score outliers, then divide each day by 100 and save the results.

When using BARD to classify BARD text summaries, the responses consisted of several different text patterns. However, there was more consistency with the output, and four regex patterns were used to parse all but 23 days. Of these, 18 consisted of sub-headings which were manually parsed, and another two had objective statements rather than numeric scores. These were re-submitted to BARD, and the numeric sentiments recorded. Interestingly, BARD refused to answer for the remaining three days, and responded with: “I’m just a language model and I can’t do that”, and “I’m a text-based AI, and that is outside of my capabilities”. By trial-and-error of resubmitting these queries to BARD, and removing a sentence each time, we were able to identify the specific sentences within each text that caused BARD to refuse our request. These were:

I. 2002-07-19: This has made it more expensive for people to borrow money to buy homes.

II. 2003-03-13: The low Treasury yields are good news for mortgage borrowers, as it makes it cheaper to borrow money to buy a home.

III. 2018-05-11: Gold is priced in dollars, so a stronger dollar makes it more expensive for investors who use other currencies to buy gold.

When asking BARD to elaborate on these sentences individually, it refused. However, when the latter half of sentences II and III were excluded, BARD responded with a discussion. It is not clear why these specific sentences resulted in BARD refusing to answer, but may be due to internal ethical/safety limits preventing a response text that could be interpreted as positioning or offering guidance or financial advice. To obtain sentiment scores for these three days, we removed the problem sentences from the prompt text, resubmitted the prompt summary to BARD, and saved the numerical response scores. We confirmed there were no sentiment score outliers, divide results by 100, and save the results.

B.3 LLM Text Sentiment Summary

We present summary statistics of sentiment measures for both text universes in Table 16. In addition, we compare daily sentiments with contemporaneous daily % returns of the S&P 500 Index, with Pearson correlation coefficients between sentiment measures and contemporaneous daily returns for both text universes presented in Figures 8 and 9.

Panel A: Summary statistics of ChatGPT text sentiments

	LM		VA		FB		TR		GP		BA	
	+	-	+	-	+	-	+	-	+	-	+	-
Mean	0.019	0.033	0.109	0.067	-0.134	1.259	0.414	-1.092	0.505	0.185	0.481	0.149
Median	0.018	0.031	0.107	0.062	-0.354	1.826	0.527	-1.256	0.605	0.090	0.500	0.100
P25	0.010	0.022	0.081	0.040	-1.413	-0.246	-0.707	-2.145	0.300	0.040	0.250	0.100
P75	0.026	0.042	0.136	0.089	1.115	2.740	1.480	-0.014	0.705	0.295	0.700	0.200
Std dev	0.012	0.015	0.040	0.037	1.316	1.587	1.321	1.265	0.257	0.218	0.236	0.124
Minimum	0.000	0.000	0.007	0.000	-1.925	-1.940	-2.349	-3.497	0.010	0.010	0.010	0.010
Maximum	0.068	0.097	0.249	0.271	2.147	3.085	3.317	1.863	0.905	0.990	0.985	0.900

Panel B: Summary statistics of BARD text sentiments

	LM		VA		FB		TR		GP		BA	
	+	-	+	-	+	-	+	-	+	-	+	-
Mean	0.011	0.037	0.088	0.073	-0.713	1.820	-0.368	-0.502	0.409	0.343	0.398	0.193
Median	0.008	0.035	0.083	0.066	-1.231	2.559	-0.650	-0.313	0.333	0.300	0.300	0.150
P25	0.003	0.024	0.063	0.043	-1.681	1.061	-1.379	-1.493	0.050	0.010	0.200	0.100
P75	0.016	0.048	0.109	0.096	0.031	2.899	0.523	0.550	0.705	0.450	0.633	0.200
Std dev	0.010	0.017	0.035	0.040	1.192	1.449	1.192	1.201	0.328	0.331	0.250	0.168
Minimum	0.000	0.000	0.012	0.000	-1.955	-2.104	-2.312	-3.310	0.010	0.010	0.010	0.010
Maximum	0.067	0.127	0.260	0.254	2.114	3.085	3.130	1.766	1.000	1.000	0.980	1.000

Note: n=5,471 days for Panel A, and 5,930 for Panel B.

Table 16: Panels A and B report summary statistics of daily positive and negative sentiment measures for ChatGPT and BARD text universes separately. The six sentiment classifiers are Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP) and BARD (BA). Results rounded to 3 decimal places.

Overall, sentiments across both text universes are similar to each other, as well as those from news text summaries in Table 2. Common to both LLM text universes, negative sentiment maximums are generally greater than positive sentiments, while sentiments for the BARD text universe generally have lower means but higher variance than those from the ChatGPT text universe. This could indicate that the BARD text universe may contain a wider range of topics and themes, and hence have a wider range of positive and negative sentiments than ChatGPT text.

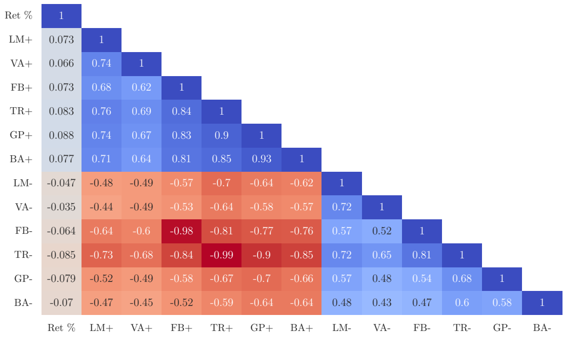


Figure 8: GPT text universe daily sentiment correlations, along with daily S&P 500 returns.

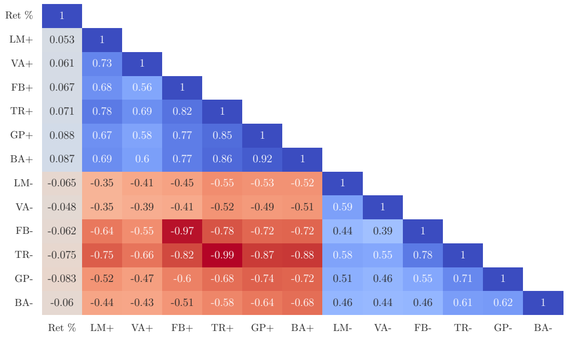


Figure 9: BARD text universe daily sentiment correlations, along with daily S&P 500 returns.

Figures 8 and 9 show similar results to that of news-based sentiments, with similar results for both LLM classified text as well as simpler dictionary and transformers. For both positive and negative sentiments across both text universes, the LLMs GP and BA have some of the strongest correlations to same-day returns, followed by FB and TR, and lastly traditional dictionary methods LM and VA. While correlations to same-day returns are weaker than news-based results of Figure 2, the results here suggest both ChatGPT and BARD recalled text is capturing similar text-based financial sentiment contained in end-of-day news summaries.

B.4 LLM Text Universe Summary

Summary statistics of ChatGPT and BARD text universes are presented in Table 17. Although the length of daily text summaries varies for both LLMs, they are generally similar. Although they are slighlty longer than They are , with BARD recalling on average 10-20% more text than ChatGPT, roughly 50-70 more words per day. Figure 10 shows the 30-day moving average of daily response length over time.[36]

[36] Almost all interaction with ChatGPT and BARD occurred over April-May 2023, with a late addendum in early August 2023. For BARD, both text extraction and sentiment classification was performed after Google updated the backend of BARD to PaLM 2 on the 10th May, Ghahramani, Z. (2023). *Introducing PALM 2*. Google. https://blog.google/technology/ai/google-palm-2-ai-large-language-model/. Accessed: 15 May 2023. Extracting daily summary text from ChatGPT took roughly 2.5 days, while BARD took 5 days. ChatGPT text was obtained without issue, however BARD sporadically timed out or responded with an error. These issues were likely due to the unofficial method of query submission to BARD, and was overcome by waiting several seconds before re-submitting the query.

	ChatGPT 1-1-2000 to 30-9-2021		BARD 1-1-2000 to 31-7-2023	
	Characters	Words	Characters	Words
Total	9,440,941	1,554,043	12,630,641	2,077,097
Mean	1,725	284	2,119	348
Std dev	190	32	484	79
Minimum	1,027	173	1,237	198
Maximum	2,584	429	4,570	756
No. days	5,472		5,962	

Table 17: Summary statistics for ChatGPT and BARD text universes.

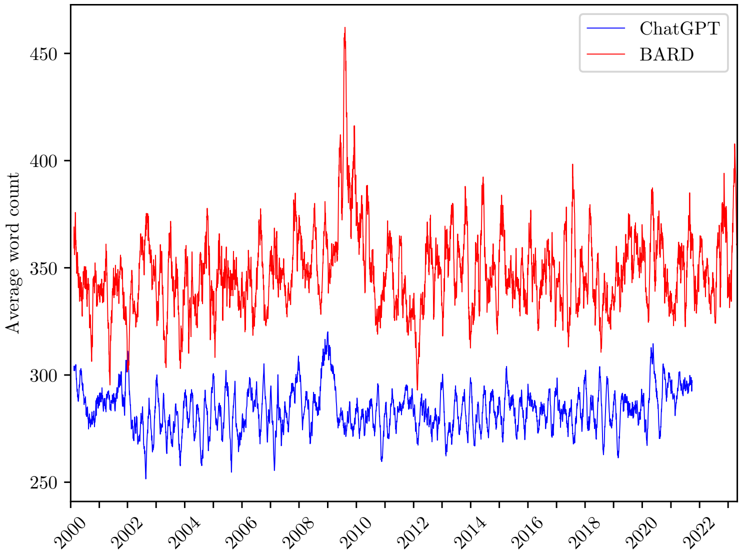


Figure 10: 30-day moving average word counts for ChatGPT and BARD text universes from 1-1-2000 onwards. ChatGPT sample ends at the knowledge cutoff date 30-9-2021. BARD sample ends 31-7-2023.

Generally, the response lengths for both ChatGPT and BARD are consistent over time. Interestingly, there is a large increase in average word count for BARD around the latter half of the GFC in 2009-2010. ChatGPT does not exhibit this same spike, but rather has a comparatively smaller spike several months earlier, in the latter half of 2008 roughly around the collapse of Lehman Brothers.

There is also a reduction in the average volume of text for both LLMs at the start of 2012, coinciding with the peak of the Greek government-debt crisis. This could suggest financial news attention, or the LLMs exposure to news, was drawn away from U.S. based markets and towards Europe at the time, as the Greek economy was close to default. There is also a small spike in the length of recalled text in early 2020 coinciding with the onset of COVID-19. Finally, the standard deviation of BARD response length is more than double that of ChatGPT, possibly hinting at more flexible and dynamic responses.

B.5 In-sample Daily Results - ChatGPT Knowledge Cutoff

Table 18 presents in-sample OLS results of Equation (1) for both text universes up to the ChatGPT knowledge cutoff date 30th September 2021. Results show sentiments for BARD text have higher adj. R^2 values, and larger AIC and BIC values than the same sentiment models for ChatGPT text.

Panel A: ChatGPT Text Universe 1-1-2000 to 30-9-2021

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.0175 *	-0.0059 *	-0.0002 ***	-0.0002 ***	-0.0016 ***	-0.0021 ***
R_t	-0.1079 ***	-0.1088 ***	-0.1087 ***	-0.1091 ***	-0.1095 ***	-0.1099 ***	-0.1093 ***
VIX_t	0.0082	0.0095	0.0093	0.0102	0.0106	0.0116 *	0.0114 *
$Tedrate_t$	-0.0014	-0.0013	-0.0013	-0.0013	-0.0012	-0.0012	-0.0012
$Termspread_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
$BAA10Y_t$	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
Intercept	0.0002	0	-0.0005	-0.0001	-0.0007	-0.0012	-0.0013
R-squared	0.0153	0.0162	0.0162	0.0171	0.0171	0.0179	0.0176
Adj. R-squared	0.0144	0.0151	0.0151	0.016	0.016	0.0168	0.0165
AIC	-32,573	-32,576	-32,576	-32,581	-32,581	-32,585	-32,584
BIC	-32,527	-32,523	-32,523	-32,528	-32,528	-32,532	-32,531
n				5,470			

Panel B: BARD Text Universe 1-1-2000 to 30-9-2021

Variables	Control	LM	VA	FB	TR	GP	BA
$Sent_t$	-	-0.0263 ***	-0.0086 ***	-0.0002 ***	-0.0003 ***	-0.0013 ***	-0.0019 ***
R_t	-0.1079 ***	-0.1099 ***	-0.1091 ***	-0.1094 ***	-0.1096 ***	-0.1102 ***	-0.1096 ***
VIX_t	0.0082	0.0092	0.0097	0.009	0.0096	0.0107 *	0.0105
$Tedrate_t$	-0.0014	-0.001	-0.0012	-0.0012	-0.0011	-0.0009	-0.0008
$Termspread_t$	-0.0002	-0.0002	-0.0002	-0.0003	-0.0003	-0.0003	-0.0003
$BAA10Y_t$	-0.0002	-0.0003	-0.0003	-0.0003	-0.0003	-0.0005	-0.0004
Intercept	0.0002	0.0007	-0.0002	0.0007	0.0002	0.0001	-0.0003
R-squared	0.0153	0.0175	0.0171	0.0169	0.0177	0.0187	0.0183
Adj. R-squared	0.0144	0.0164	0.016	0.0158	0.0166	0.0176	0.0172
AIC	-32,573	-32,583	-32,581	-32,580	-32,584	-32,590	-32,587
BIC	-32,527	-32,530	-32,528	-32,527	-32,531	-32,537	-32,535
n				5,470			

Table 18: In-sample daily sentiments and next day market returns up to 30th September 2021.
This table reports regression constants β from the following model: $R_{t+1} = \alpha + \beta_1 Sent_{t,C} + \beta_2 R_t + \beta_3 VIX_t + \beta_4 Tedrate_t + \beta_5 Termspread_t + \beta_6 BAA10Y_t + \varepsilon_{t+1}$, where R_{t+1} is the next day % change of the S&P 500, and $Sent_{t,C}$ is the overall daily pessimism sentiment measure for classifier C, one of the six sentiment classifiers: Loughran & McDonald (LM), VADER (VA), FinBERT (FB) TwitterRoBERTa (TR), ChatGPT (GP) and BARD (BA). R_t is the daily % change of the S&P 500, VIX_t is the daily CBOE Volatility Index/100, and $Tedrate$, $Termspread$ and $BAA10Y$ are measures of daily short, medium, and long-term credit risks, as detailed in Section 3.1. The sample period covers 1st January 2000 to 30th September 2021 for both text universes. Robust standard errors are used. $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.