EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Fabian Wagner, Julian Metzler

When the crowd speaks: AI-based retail investor sentiment indicator with Reddit data

No 3276

Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

Retail investors increasingly discuss markets in real time on social media, yet these discussions remain difficult to measure systematically. This paper introduces the Reddit Retail Investor Sentiment Indicator (R-RISI), a high-frequency measure of retail investor sentiment based on Reddit posts from major investing and crypto-related subreddits between 2015 and April 2026. Using ChatGPT 5.1, we classify posts by sentiment and topic and show that large language models provide more accurate labels for informal social media language than widely used finance-specific models such as FinBERT, which tends to over-assign neutral sentiment. R-RISI is constructed by aggregating daily positive and negative posts, weighting them by user engagement and standardising the resulting series using a rolling one-year z-score methodology. The indicator can be flexibly built for broad asset classes, portfolios, individual securities or thematic groups, allowing sentiment to be tracked at a much higher granularity than traditional retail investor indicators. We show how sentiment and dominant discussion topics evolve over time within each of the extracted subreddits. R-RISI closely reflects major market developments and aligns with established sentiment indicators while providing higher frequency and more timely signals. Empirically, changes in R-RISI contain statistically significant short-term information for market prices. This relationship is particularly relevant during periods of market stress: changes in R-RISI matter more for next-day S&P 500 returns when markets are in decline. The indicator therefore offers a timely and granular tool for analysing retail investor behaviour and monitoring sentiment-driven market dynamics in an environment of growing retail investor participation. A regularly updated version of the indicator is available through an accompanying R-RISI online dashboard.

Keywords: retail investors, behavioural finance, AI/LLM-based text classification, sentiment analysis, Reddit, social media, alternative data

JEL Codes: F1, G1, G4, G5

Non-technical summary

We introduce a high-frequency indicator of retail investor sentiment from social media: the Reddit Retail Investor Sentiment Indicator (R-RISI). It transforms unstructured Reddit posts into a daily measure of investor optimism and pessimism using artificial intelligence.

We analyse over eight million Reddit posts (titles and bodies) from major investing- and crypto-related Reddit communities between 2015 and April 2026. Each post is classified by ChatGPT-5.1 by sentiment (positive, negative, or neutral) and topic (equities, funds, macro, options, commodities, geopolitics, tariffs, crypto, or other). Posts labelled “other” are excluded to focus solely on investment discussions. For comparison, we apply FinBERT, a common financial sentiment model, but find that it performs poorly on informal social-media language, over-assigning neutrality and missing irony, slang, and humour. ChatGPT-5.1 yields more meaningful sentiment classifications and is used for the final indicator.

R-RISI aggregates daily positive and negative posts, weighting them by user engagement (log of upvotes plus comments), so posts drawing more attention influence the indicator more. Both sentiment series are standardised using a rolling one-year z-score to ensure comparability over time. Net sentiment is defined as the difference between standardised positive and negative components, providing a scale-free measure of unusual optimism or pessimism.

This framework produces a flexible and granular indicator that can be computed not only for broad equity or crypto markets, but also for specific sectors, portfolios, or individual assets. R-RISI can also be broken down by topic, allowing users to track sentiment related to macroeconomic issues, geopolitical risks, or other narrative drivers of market behaviour.

We show that R-RISI closely mirrors market developments and reacts swiftly to global events and stress episodes. It correlates strongly with equity prices and established sentiment measures, while offering higher-frequency and faster signals. Changes in R-RISI have statistically significant short-term price effects, particularly during market stress, suggesting that online discussions may reflect actual retail trading. Reddit discussions also appear topical and timely.

R-RISI provides a novel tool for monitoring retail investor sentiment in near real time. For researchers, it offers a scalable and transparent methodology to analyse sentiment-driven market dynamics across assets and themes. For policymakers and financial stability authorities, it delivers a timely indicator of shifts in market mood and speculative dynamics among retail investors, helping to identify emerging risks in financial markets.

1 Introduction

Once dismissed as internet slang on Reddit's r/wallstreetbets, expressions such as "to the moon," "diamond hands," and "YOLO trades" now symbolize a new era of retail investing. The GameStop rally showed how individual investors coordinating online can amplify sentiment and generate market-wide effects. In this paper, we introduce a method to extract a standardized daily Reddit Retail Investor Sentiment Indicator (R-RISI) from investing-related Reddit posts, interpreted by the large language model ChatGPT-5.1. We show that R-RISI tracks market prices, captures events at high frequency and helps explain market movements. Empirically, we find that retail sentiment has a short-term effect on market prices within one trading day.

Over the past decade, retail participation has transformed from a niche phenomenon into a central feature of modern financial markets, driven by accessible apps, zero-commission trading, and social media. In the U.S., retail investors' share of equity trading rose from 10% in 2010 to nearly 25% during the pandemic, before easing to around 20% by mid-2025 and 17% in Q1 2026. Retail investors' inflows reached a decade-high USD 4.7 billion in April 2025 amid the tariff announcements (Reuters, 2025). Because retail investors collectively can amplify market swings and affect financial stability, they have attracted growing academic and policy attention. Yet timely measures of retail activity remain scarce.

Traditional survey-based indicators, such as the AAII Sentiment Index or the Charles Schwab Trading Activity Index™ (STAX), offer useful benchmarks but are limited in frequency, coverage, and responsiveness. Meanwhile, the growing activity within online trading communities on social media platforms such as Reddit, Twitter, and StockTwits offers an abundant and largely untapped data source that reflects the collective mood of more speculative retail investors. Reddit, in particular, stands out due to its large user base, topic-specific subreddit structure, rich discussion threads and freely accessible data. This paper facilitates this high-frequency and highly granular datasource to gain new insights into retail investor behaviour.

We develop a high-frequency sentiment indicator that quantifies retail investor mood using millions of Reddit posts from major investing and crypto-related subreddits between 2015 and end of April 2026. Employing advanced natural language processing (FinBert, OpenAI's GPT-4o and their GPT-5.1), we derive a daily sentiment measure and compare it with global equity prices and established retail investor indicators. We also apply AI based topic labelling to identify dominant discussion themes. This approach captures both sentiment and content, providing deeper insight into retail investor positioning over time.

Our findings show that Reddit-based retail investor sentiment closely co-moves with equity markets, with correlations rising during stress, suggesting some patterns of procyclical dynamics in retail investor mood. R-RISI tracks established sentiment indices but offers higher frequency, lower lag and potentially less measurement bias than traditional surveys. Our empirical analysis shows that at a daily frequency, R-RISI drives part of market price dynamics. An increase in daily retail investor sentiment, even when controlling for price spillovers from previous days into current sentiment, is associated with higher market prices on the next day. We interpret the positive and significant coefficient as evidence that sentiment expressed on Reddit partly translates into retail investor behaviour and short-lived market effects. We further show that R-RISI has a stronger representativeness of retail investor activity during market downturns as opposed to recovery or calm periods, providing a basis for future research on behavioural dynamics and market-based financial stability.

Our novel indicator provides a scalable, timely proxy for retail investor sentiment. It enables granular analysis of sentiment-driven market movements, gives policymakers a real-time tool to monitor household discussions, and offers a dynamic signal of crowd mood. Overall, the results highlight the growing role of retail investors and the value of social media-based sentiment indicators for behavioural finance, market microstructure and financial stability.

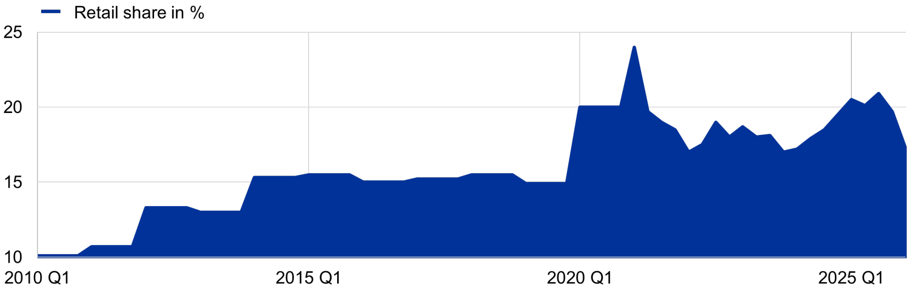
The remainder of this paper is structured as follows: Section 2 provides an overview on the recent literature. Section 3 introduces the Reddit data underlying the R-RISI as well as market data, and Section 4 explains the methodology for calculating R-RISI using AI tools. Section 5 describes the results. We contrast R-RISI against other retail activity measures, as well as equity prices. Finally, Section 6 tests the effect of R-RISI on price indices. Section 7 concludes.

2 Literature Review

Academic finance has traditionally focused on institutional investors, but technological advances have increased the role of retail investors. As direct trading data are scarce, research increasingly relies on social media and online forums to proxy retail sentiment and expectations. A large literature shows that sentiment influences market outcomes in the short term, and text based studies using news and social media have grown rapidly since the early 2000s.

Early work focused on aggregate sentiment measures and their asset-pricing effects. Baker

Figure 1: Share of Retail Investors in US Equity Trading Volume



Note: The retail trading share is derived from Bloomberg Intelligence (BI MKTS) estimates of the composition of US equity trading volume. The series reflects quarterly assessments of the retail investor segment's share in total trading. Data is shown from Q1 2010 to Q1 2026 and has the unit percent. Source: Bloomberg Finance L.P.

and Wurgler (2006, 2007) constructed composite sentiment indices and showed that sentiment predicts the cross-section of stock returns, especially for hard-to-arbitrage stocks. Tetlock (2007) found that pessimistic Wall Street Journal news predict falling stock prices and higher trading volume, establishing a direct link between media tone and market outcomes. García (2013) extended this work by showing that news' impact differs between recessions and boom periods.

Later studies refined text-based sentiment measures and extended them to firm-level and high-frequency settings. Tetlock et al. (2008) showed that negative words in firm news predict earnings and returns, while Antweiler and Frank (2004) found that message board activity and sentiment are strongly linked to trading volume and volatility, though return predictability is economically modest. These studies highlight two recurring features of sentiment effects: timeliness and short-lived impact.

From the mid 2000s, research increasingly examined sentiment as a predictor of volatility and uncertainty. Large news-based studies showed that sentiment and disagreement improve volatility forecasts and capture shifts in investor beliefs (Engelberg & Parsons, 2011). Joseph et al. (2011) found that online ticker searches affect abnormal returns and trading volume. Similarly, Da et al. (2015) introduced the Google Search Volume Index as a measure of investor attention and showed that it predicts returns and volatility, supporting the view that text-based indicators proxy attention and information demand rather than long-term risk premia. Based on Bloomberg sentiment data, Seetharam and Nyakurukwa (2024) find that information mainly flows from social media to news for most stocks, whereas the reverse holds for few firms.

The rise of social media marked a turning point in the literature. Platforms such as Twitter, StockTwits, and Reddit provide real-time user-generated content that reflects retail investor sentiment. Twitter sentiment, for instance, shows predictive power for short-horizon returns and volatility, though effects fade quickly (Bollen et al., 2011). More recent quasi-experimental evidence shows that social-media sentiment has causal short-term effects on stock prices (Wang et al., 2022). Research on meme stock episodes further shows that concentrated retail sentiment can trigger sharp but temporary price movements and volatility spikes (Barber et al., 2022). Likewise, Semenova and Winkler (2025) analysed WallStreetBets text data to show that peer driven investor sentiment affects retail trading and contributes to bubble like price run ups.

After Twitter's rebranding to X, data access restrictions have reduced its usefulness for large-scale research, making Reddit data more attractive. Recent studies link Reddit activity and sentiment to market prices and short-term movements, particularly during the GameStop episode and for speculative assets such as Bitcoin. However, evidence remains limited to short samples and specific events (e.g. Affuso & Lahtinen, 2019; Li & Li, 2026; Long et al., 2022, 2025). Münster et al. (2024) show that both traditional media and social media (proxied by Reddit posts) affect retail investors' trading decisions, with social media exerting a substantially stronger influence. Liu and Son (2024) review methods for aggregating investor sentiment from social media and highlight key challenges in textual data, including neutral sentiment and the strong dependence of index reliability on construction choices.

A growing body of publications, mainly conference papers and SSRN research applies more advanced transformer-based financial language models to social media text to extract sentiment and evaluate predictability (e.g. Chen et al., 2026; Huang & Shum, 2025). Applications have shifted from dictionary-based methods to transformer models such as VADER and FinBERT. Large language models classify sentiment and topics from unstructured text using semantic representations. FinBERT adapts Google's BERT to financial news but may perform poorly on informal social media language with slang and irony (Araci, 2019). Investor sentiment measures perform better when using finance-specific dictionary models (Ballinari & Behrendt, 2021). Although Rodríguez-Ibáñez et al. (2023) find that pre-trained LLMs outperform generative models such as GPT-3 in cost effectiveness, newer and more efficient generative models such as ChatGPT-4o enable new approaches to analysing social media text at scale, making large-scale applications substantially cheaper. Generative AI also has limitations. While transformer models produce consistent sentiment labels, generative LLMs are black boxes and may yield slight output variation when promted again. Mun and Kim (2025) show that prompt design and settings affect accuracy. We advance sentiment analysis by applying the generative model ChatGPT-4o and GPT-5.1 to extract sentiment, content and tickers from Reddit posts at scale. Compared to FinBERT, ChatGPT models better captures nuanced tone, while FinBERT often defaults to neutral, especially with slang. Baklanova (2025) and Nguyen et al. (2025) follow a similar approach focusing on Bitcoin, analysing Reddit and Twitter sentiment, respectively. The former uses Flair and VADER, while the latter compares several LLMs, including advanced GPT models and finds them outperforming transformer based models (data: 2018-2023).

Existing studies highlight retail sentiment and social media but focus mainly on short horizons and events such as GameStop. To our knowledge, no long-horizon, high-frequency Reddit-based sentiment indicator exists that is flexible across indices, asset classes and topics, while using AI to extract sentiment and themes without prior domain training. We fill this gap by constructing the Reddit Retail Investor Sentiment Indicator (R-RISI) with ChatGPT 5.1, using millions of posts from major investing, crypto, macro and commodities subreddits over more than a decade. R-RISI provides a new tool to analyse the link between retail sentiment, market dynamics and financial stability.

3 Data

This paper draws on three data sources: Reddit to capture retail investor sentiment (Section 3.1), market data, and complementary alternative sentiment indicators (Section 3.2).

3.1 Reddit data

Reddit is one of the world's largest discussion platforms, organised into thematic communities called subreddits. It has become a central venue for retail investors to exchange views and discuss market developments. Since 2020, Reddit – especially r/wallstreetbets, a subreddit – has played a prominent role as a source of information on trading strategies for retail investors. Users are mostly individual investors who discuss equities, options, cryptocurrencies, and related topics. Posts are made under anonymous usernames, allowing open and candid discussion. Reddit thus provides an unfiltered, high-frequency view of retail investor sentiment and attention dynamics.

The raw Reddit data are obtained from Reddit Academic Torrents (see RaiderBDev, 2026), a publicly available archive distributed via the Academic Torrents platform. This peer-to-peer (P2P) system, based on the BitTorrent protocol, enables efficient sharing of large research datasets. The dataset is provided by the Pushshift project and is updated typically mid-month. As part of ongoing work, we incorporate new Reddit data and further develop our R-RISI online dashboard.

We include the largest Reddit communities related to investing, cryptocurrencies, macro and commodities, selected by topical relevance and popularity, i.e. subscriber count. For instance, r/wallstreetbets captures highly active retail trading discussions, while r/CryptoCurrency focuses on digital assets and blockchain topics. A detailed overview of all included subreddits is provided in Table 1. To streamline the analysis and manage computational resources, we restrict the dataset to Reddit post titles and their respecitive text bodies (submissions in the Pushshift data) and exclude comments. This reduces file size and LLM labeling costs, while we still capture community engagement through post-level metrics as discussed later. The raw dataset contains key metadata for each post, including unique identifiers, timestamps, titles, and engagement metrics. Variables are shown in Table 2, and a comprehensive documentation of the dataset and its collection methodology is provided by Baumgartner et al. (2020).

Table 1: Overview of Popular Reddit Communities Focused on Investing and Crypto

Subreddit	**Description**	**Main Topics**	**Category**	**Members**
r/wallstreetbets	High-risk, meme-heavy retail community	Meme stocks, option or YOLO trades	Investing	19.9 M
r/stocks	Discussions on individual stocks and news	Stock analysis, earnings, valuation, news	Investing	9.2 M
r/StockMarket	Focused on broader market and trading	Market commentary and trading ideas	Investing	4.0 M
r/investing	General investing discussion community	Long-term investing, portfolios	Investing	3.3 M
r/options	Dedicated to option trading	Greeks, risk mngmt	Investing	1.4 M
r/CryptoCurrency	General discussion on cryptocurrencies	Crypto markets, coins, exchanges	Crypto	10.1 M
r/Bitcoin	Bitcoin-specific community	Bitcoin price development	Crypto	8.1 M
r/Ethereum	Ethereum-focused community	Smart contracts, dApps, DeFi	Crypto	3.7 M
r/CryptoMarkets	Trading-focused forum for digital assets	Crypto related posts	Crypto	2.0 M
r/Economics	Economics discussion community	Economics, policy, macro, labour, trade	Macro	5.7 M
r/geopolitics	Geopolitical news discussions	International relations, conflicts, politics	Macro	770.4 K
r/Gold	Gold-focused precious metals community	Physical gold, coins, bars, buying/selling	Commodity	377.3 K
r/Wallstreetsilver	Silver-focused community	Physical silver, buying/selling	Commodity	286.6 K
r/energy	Energy news discussions	Renewables, nuclear, fossil fuels, storage, energy policy	Commodity	273.7 K

Note: This table presents our selection of popular Reddit subreddits related to investing, crypto, macro and commodities. "Members" indicates the number of subscribers as reported by Reddit as of April 2026; these figures are expected to grow further with time.

Table 2: Overview of Used Variables from the Raw Reddit Post Dataset

Field	Description
id	Submission's identifier, e.g., "l72w5z" (String)
created_utc	UNIX timestamp of the post's creation, e.g., *1611854398* (Integer)
subreddit	Name of the subreddit of the post, e.g., *"WallStreetBets"* (String)
title	Title of the post, e.g., *"Is $SNDL gonna go to the moon???"* (String)
selftext	Body of the post (String)
num_comments	Number of comments associated with a submission, e.g., *358* (Integer)
score	Number of upvotes minus the number of downvotes, e.g. *1256* (Integer)
author	Account name of the poster (String)
url	Post link, e.g. *https://www.reddit.com/r/wallstreetbets/comments/l72w5z/is_sndl_gonna_go_to_the_moon/* (String)

Note: The table presents the variables used in this study, its data type, and unit, and provides an example. For a documentation and additional fields see Baumgartner et al. (2020). During the analyses variables were added, e.g. sentiment and topic labels of which the construction is described in Section 4.1.

To ensure data quality and reliability for the later sentiment and topic labelling (Section 4.1), our initial FinBERT and ChatGPT-4o classifications used cleaned Reddit post titles only. We removed missing, duplicate and non-informative titles, such as emoji-only or very short titles, retained the most recent post for identical subreddit–author–title combinations to preserve updated engagement metrics, cleaned tags, links and HTML fragments, converted emojis into descriptive text, normalised text, and discarded title-length outliers above the 98th percentile. However, after observing a moderate share of neutral sentiment and residual-topic labels, we manually checked the data and found that titles alone often omit relevant context from the post body. We therefore relabelled the dataset using both titles and selftext as input, based on ChatGPT-5.1, the latest model available at our institution. We used the raw combined text, as the reasoning model can process noisier inputs and use the additional context more effectively. UNIX timestamps are reported in UTC and were transformed into date–hour–second format. We removed posts labeled "[deleted by user]" and "[Removed by moderator]", the only outlier duplicate title occurrences.

While Reddit data provide a rich, unfiltered view of retail investors' discussions and reactions to market developments, they remain underused in large-scale sentiment analyses, adding to the novelty of this paper. Reddit offers a valuable source for detecting sentiment shifts but the data comes with limitations. In terms of representativeness, Reddit users are not a random sample of the overall retail investor population. The platform mainly attracts younger, male, and technologically inclined individuals, largely based in the United States (see e.g. Aubin & Liedke, 2025; Hernandez, 2025), which may bias sentiment toward speculative or risk-seeking views. Evidence also suggests that Reddit users favor higher risk, lottery type trading strategies rather than mean variance efficient portfolios (Reichenbach & Walther, 2023). Most discussions are in English, and regional subreddits in other languages were excluded. Post titles capture only part of the discussion and may include sarcasm, humor, or coordinated narratives that distort tone. Many posts also contain images or screenshots, which were not considered for the sentiment calculation. Posting behaviour is additionally shaped by subreddit-specific moderation rules. Large communities such as r/wallstreetbets or r/stocks employ automated filters that remove specific types of content, for instance cryptocurrency-related posts in r/wallstreetbets are automatically deleted. Extreme insults or hate speech are also filtered out by moderation bots. Moreover, several subreddits require users to have accumulated a minimum amount of "karma", i.e upvotes, before they are allowed to post or comment. For example, r/wallstreetbets requires 50 karma points for comments and 100 for submissions (based on self-testing). These rules restrict participation by newer or less-established users and can influence which sentiments and narratives become visible in the dataset. It further remains unclear to what extent expressed sentiment translates into actual trading. Discussions can also reflect herd behaviour, amplifying narratives beyond their market relevance. Nonetheless, Reddit data are largely objective, as content arises organically without survey framing or researcher interference, capturing authentic peer-to-peer communication and the Reddit data is available at no costs. Although content quality may vary over time due to moderation, community shifts, or bot activity, Reddit provides a unique high-frequency perspective on particularly active, sentiment-driven market participants.

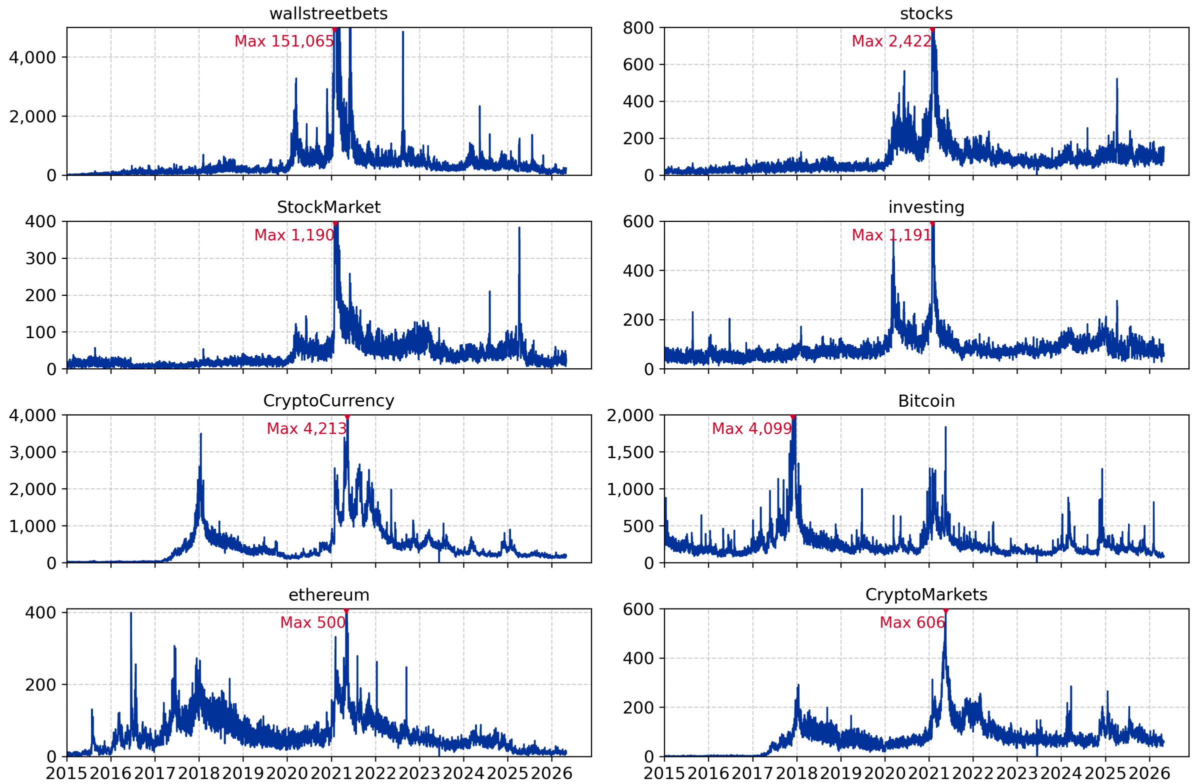
Our dataset, covering eight investing- and crypto-related subreddits, as well as selected commodity- and macro-related subreddits, comprises over 8.49 million Reddit posts between January 2015 and April 2026. Detailed descriptive statistics on post counts and engagement metrics are reported in Appendix B. For r/wallstreetbets, for instance, the median post volume is 263 per day over the full sample. Post titles contain on average 8.96 words, receive 34.14 comments, and achieve a mean score of 105.90. Excluding 2021 slightly lowers these values, reflecting the extreme posting activity during the meme-stock episode. Among all forums, r/wallstreetbets dominates overall activity, far exceeding subreddits such as r/stocks or r/investing in cumulative post counts (see Appendix B). In the paper, we focus mainly on investing- and crypto subreddits.

As shown in Figure 2, daily posting activity fluctuates over time, with distinct spikes around market-moving events. Activity rose sharply during the early-2021 retail trading frenzy and, to a lesser extent, during major crypto booms. Posting behaviour is broadly balanced across weekdays, though weekends show about 5–7 percentage points lower activity (Figure 3, Panel A).

Intraday activity follows a clear pattern: it increases through the morning, peaks between 13:00 and 16:00 UTC (9 a.m.–12 p.m. New York time), and declines in the evening (Figure 3, Panel B), reflecting discussion peaks around U.S. market openings. The increased activity during U.S. market opening hours indicates the timeliness of the information that can be extracted from Reddit data. Engagement on the platform remains high until the U.S. stock market closes, and then sharply drops within the hours following market closure. We interpret this as a sign of immediacy of Reddit discussions.

Figure 4 shows the evolution of unique authors: daily active accounts remained moderate until late 2019 but surged in early 2021, reaching nearly 100,000 in a single day. By 2026, the cumulative number of distinct authors exceeded 2 million, illustrating the growing breadth of retail investor participation in online forums during our sample period.

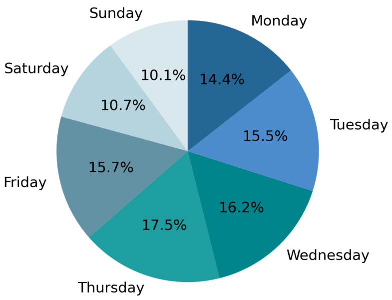
Figure 2: Daily Posting Activity Across Major Investing and Crypto-related Subreddits



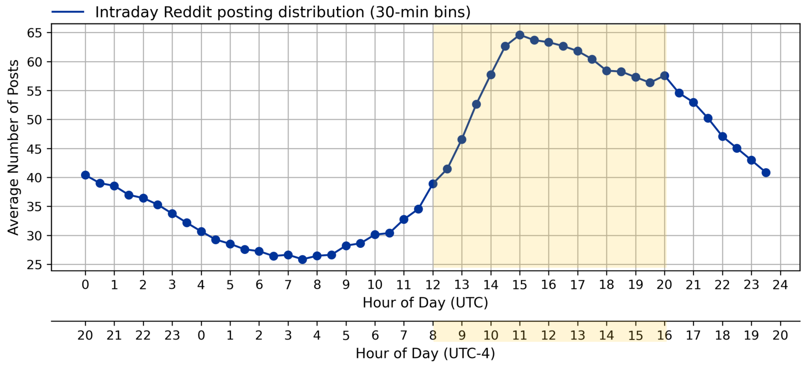
Note: The figure shows the daily number of Reddit posts between January 2015 and April 2026 across eight major investing and crypto subreddits. Posting activity shows pronounced spikes around market-moving events, such as during the Covid pandemic. The y-axis is truncated and the respective maximum value is indicated in red. Post counts are aggregated by day and subreddit based on post creation timestamps. Posting volume for the additional subreddits can be found in Appendix B.

Figure 3: Reddit Posting Activity by Weekday and Intraday for an Average Day

Panel A: Posting Activity by Weekday

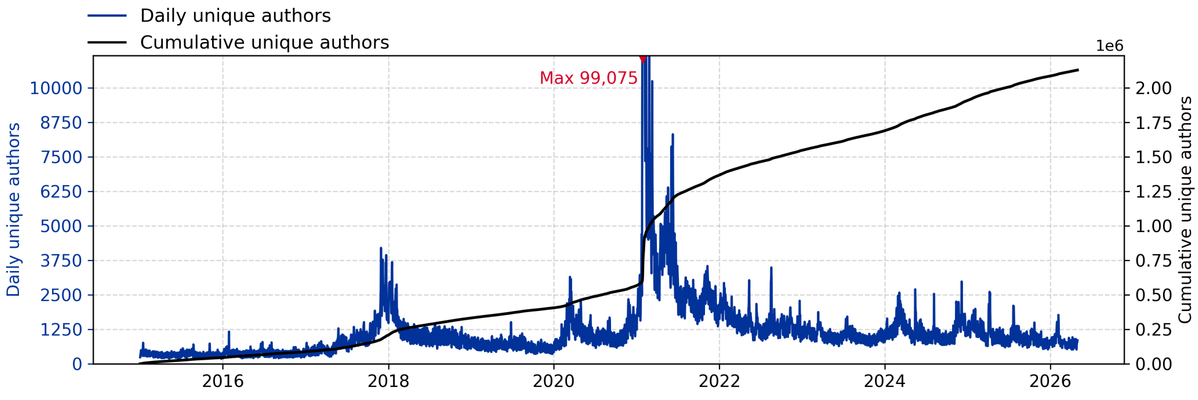


Panel B: Intraday Posting Distribution (Average Day)



Note: Panel A shows the Reddit post activity by weekday, based on the full dataset period. Panel B illustrates the intraday posting distribution in 30-minute bins, averaged across all days in the sample. Timestamps are recorded in Coordinated Universal Time (UTC); the secondary x-axis converts these to Eastern Daylight Time (UTC−4), showing U.S. market hours in New York City. The yellow area marks when exchanges are open there.

Figure 4: Daily and Cumulative Number of Unique Reddit Authors in the Selected Subreddits



Note: The figure reports the daily number of unique authors posting the above mentioned subreddits (left axis) and the cumulative total number of unique authors over time (right axis). Author activity peaked during the early-2021 meme-stock episode, when nearly 100,000 individual accounts posted within a single day.

3.2 Market Data, Sentiment Benchmarks and Data Merging

Market data are obtained from Bloomberg Finance L.P. We collect daily closing prices for major equity indices, major crypto assets and equities frequently discussed on Reddit. These data are used to benchmark the Reddit-based sentiment indicator against market dynamics.

To validate our indicator, we also use available retail investor sentiment benchmarks. The AAII Sentiment Survey, published weekly by the American Association of Individual Investors, reports the share of U.S. respondents who are bullish, bearish or neutral over a six-month horizon (see AAII, 2026) and is publicly available via the AAII website. As a survey-based measure, it may be affected by self-selection and non-response bias, and responses are constrained by predefined questions, unlike the more spontaneous discussions observed on Reddit.

A behaviour-based benchmark is the Charles Schwab Trading Activity Index (STAX), published monthly (see Charles Schwab, 2026). STAX reflects actual portfolio positioning and trading activity of millions of U.S. retail accounts at Schwab, offering a more representative view of broader and typically wealthier retail investors. Unlike survey-based indicators, STAX captures realised trading behaviour rather than stated intentions.

Data are aligned at the daily level. Since market data are available as daily closing prices and Reddit data at the post (tick) level, Reddit activity is aggregated by post date and merged with daily market data. As a result, all posts on a given day share the same closing market values, so intraday variation is not captured. This simplification is intentional, as R-RISI is constructed at daily frequency to limit sensitivity to short-term fluctuations. Because Reddit activity also occurs on weekends and public holidays, market data are forward filled on non-trading days to maintain continuity of the Reddit series. The AAII and STAX benchmarks are merged on their original reporting dates and compared at their respective weekly and monthly frequencies.

4 Methodology

This section outlines the methodological framework for transforming Reddit posts into a quantitative retail investor sentiment indicator. The process involves two steps: extracting post-level sentiment and topic information using advanced language models (Section 4.1) and aggregating these classifications into a standardized daily indicator, our R-RISI (Section 4.2). Together, these steps convert unstructured social media text into a structured measure of shifts in retail investor sentiment over time.

4.1 Sentiment and Topic labeling

Sentiment analysis aims to determine the emotional tone expressed in a given text, classifying it as positive, negative, or neutral. Topic labeling identifies the main theme of the post. Applying these techniques to Reddit data enables a systematic understanding of the attitudes and discussion themes prevailing among retail investors.

We first employed FinBERT (Araci, 2019), a transformer-based language model pre-trained on financial text corpora, to classify the sentiment of Reddit post titles. FinBERT categorizes text into positive, negative, or neutral classes based on a lexicon-informed, context-sensitive embedding architecture optimized for financial news and analyst reports. While initial results were broadly consistent with expectations, the model displayed a strong tendency to assign a neutral label in ambiguous or context-dependent cases. This behavior was particularly evident when posts conveyed irony, humor, or sentiment expressed through emojis, elements that FinBERT, given its training data, cannot fully interpret.

To overcome these limitations, we subsequently employed OpenAI's ChatGPT-4o model. Each Reddit post title was submitted to the model, which we prompted to return three key variables: (i) a *sentiment* label (*positive, negative, or neutral*), (ii) a *sentiment confidence* score on a 0–1 scale representing the model's certainty in its classification, and (iii) a *topic* tag assigned to one of nine pre-specified categories relevant for market analysis: *commodities, funds, options, crypto, geopolitical, equities, macro, tariffs, and other.* These categories were selected to capture the range of subjects most relevant to retail investor discussions. The *other* category includes posts not directly related to financial topics, such as moderator announcements, community discussions, or unrelated content. Table 3 provides an overview of the topic and sentiment categories, along with illustrative example posts for each.

All large language model (LLM) prompts are reported in Appendix C. Following the initial GPT-4o labelling, we found that titles alone often provided insufficient context to extract the required information reliably. We therefore relabelled the full dataset using both Reddit post titles and text bodies, as described in Section 3.1. For this final classification, we used OpenAI's ChatGPT-5.1 model, the latest model available at our institution at the time of the analysis. In addition to the sentiment label, sentiment confidence and topic label, the model was also prompted to return a *topic confidence* score on a 0–1 scale representing the model's certainty in its topic classification and several additional fields: a short *free topic* description summarizing

Table 3: Overview of Topic and Sentiment Categories with Definitions and Illustrative Example Posts

Panel A: Topic Categories

Topic Category	Description	Example Post
Equities	General stock-market discussions, company news, trading behavior, or specific tickers.	"I made 69% in 3 months holding $HOOD" – id: 1nur1s6, wallstreetbets
Crypto	Discussions about cryptocurrencies, tokens, blockchain, stablecoins, or crypto markets.	"Tether just bought over 8800 #bitcoin worth $1 billion USD." – id: 1nuso1y, Bitcoin
Options	Posts mentioning options trading, derivatives, or related terms (calls, puts, IV, Greeks).	"Another $6K + trading morning SPY 0DTE options." – id: 1nue9a6, wallstreetbets
Funds	Mentions of ETFs, mutual funds, hedge funds, or index trackers.	"TQQQ - The perfect ETF for a strong bull market" – id: opjsgd, StockMarket
Commodities	Topics about physical commodities (oil, gold, copper, wheat, etc.).	"YOLOD $20,000 into Silver" – id: l78cnx, wallstreetbets
Macro	Economy-wide or policy-related discussions: inflation, GDP, interest rates, central bank policy.	"I see some saying the market could crash" – id: l9cx26, wallstreetbets
Geopolitical	Political or military developments affecting markets, excluding trade-related topics.	"Wow! Prepare for Monday! Isarel has just declared war on Humas/ Palestine. Israeli troops moving across the border. Bombs going off" – id: 172e99r, wallstreetbets
Tariffs	Discussions of trade barriers, duties, or tariff-related international policy.	"Trump increases tariff on Canada to 35%, White House says" – id: 1megjl8, stocks
Other	Non-financial or off-topic content (memes, jokes, personal stories).	"Looking for Broker recommendations as I am planning on leaving RH." – id: lfhsbn, stocks

Panel B: Sentiment Categories

Sentiment	Description	Example Post
Positive	Posts expressing optimism, gains, or confidence in market moves.	"$POWW blew away earnings again, and raised 2022 Rev Guidance!" – id: p5suda, StockMarket
Negative	Posts expressing losses, frustration, or pessimism.	"Dow drops 1,200 points to new low on the day after WHO declares coronavirus a global pandemic" – id: fh074w, wallstreetbets
Neutral	Informational or balanced posts without clear sentiment.	"The Signs are in The Charts" – id: m5mx12, wallstreetbets

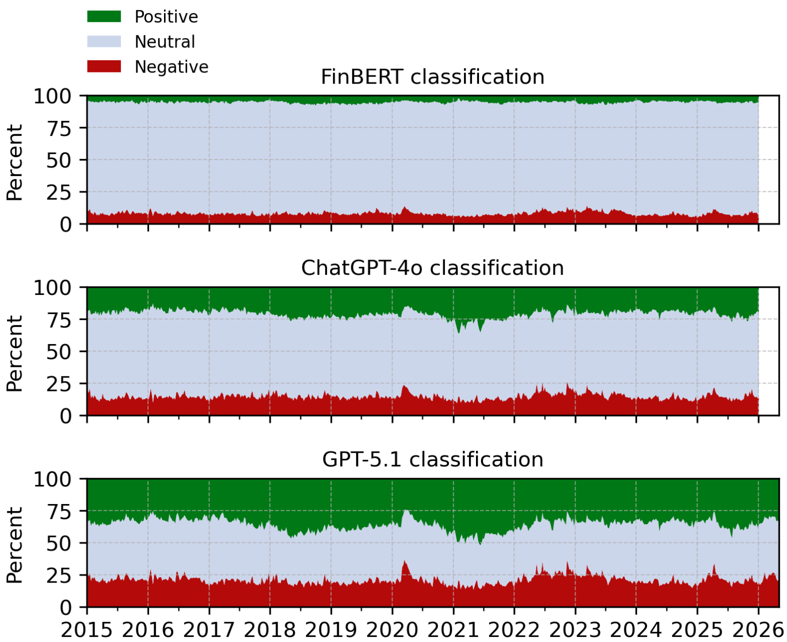
Note: Panel A lists the topic categories assigned to Reddit posts, while Panel B reports the sentiment categories. Each category is accompanied by a brief definition and an illustrative example post, including its Reddit post ID.

the post's narrative in two to three words, a list of tickers mentioned in the post *tickers*, the implied *trade direction*, the prevailing *market emotion*, the relevant *region*, and the implied *time horizon*. The ticker field records standard trading symbols where identifiable and an empty list otherwise. Ticker extraction is quite central for our application. As discussed below, R-RISI can be constructed not only for the aggregate retail-investor discussion space, but also for individual assets or portfolios of assets. Reliable ticker identification is therefore essential for asset-specific sentiment measures. We initially did not ask GPT-4o to identify tickers and instead used deterministic rules and dictionaries, for example by mapping company names to ticker symbols, detecting common crypto terms such as Bitcoin or BTC, and filtering capitalized terms following a dollar sign. However, this approach proved only partially reliable. Reddit users frequently use slang, abbreviations, misspellings and informal company references, which makes simple string matching prone to both false negatives and false positives. For example, a rule-based filter for the capital letter "F" would not reliably distinguish references to Ford Motor Company from unrelated uses of the letter. In the relabelling process we addressed this issue. The reasoning-capable GPT 5.1 LLM is well suited to this task, as it can use the surrounding text to infer whether a ticker or company is actually being referenced. The remaining auxiliary variables classify posts into broader categories, such as long, short, hedged, neutral or unclear for *trade direction*; panicked, fearful, cautious, neutral, optimistic or euphoric for *market emotion*; US, EU, Asia, Global or Other for *region*; and intraday, short-term, medium-term, long-term or unclear for *time horizon*. While these auxiliary variables provide additional information, they are more difficult to infer reliably from Reddit posts than the core sentiment and topic labels. Reddit posts are often short, informal, ambiguous or ironic, making it difficult to extract nuanced dimensions such as trade direction or investment horizon. These variables are therefore used primarily as supplementary information and should be interpreted with caution, especially in quantitative analyses. We base our final R-RISI on results provided by GPT-5.1.

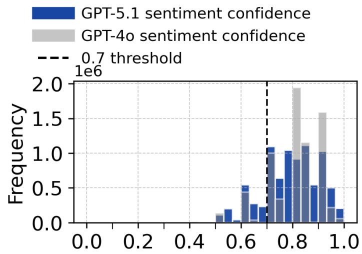
For reliability purposes, we use confidence-based filters to reduce the influence of uncertain classifications. This means that we exclude posts with a sentiment or topic confidence score of below 0.7. Figure 5 shows the number of posts by sentiment label and the corresponding distribution of sentiment and topic confidence scores before filtering. FinBERT produces a substantially more neutral-dominated classification, whereas the OpenAI LLM classifications assign sentiment with greater dispersion across positive and negative categories. This difference likely reflects the models' distinct training and usage settings: FinBERT is a domain-specific classifier fine-tuned on financial texts, which makes it conservative and prone to neutral assignments when language is ambiguous, while the LLMs can draw on broader contextual understanding and better capture colloquial, figurative and ironic expressions common on Reddit. The comparison between ChatGPT-4o and GPT-5.1 is not a direct model comparison, as both the underlying model and textual input differ (use of title versus title + text body).

Figure 5: Sentiment Labeling and Confidence Distribution

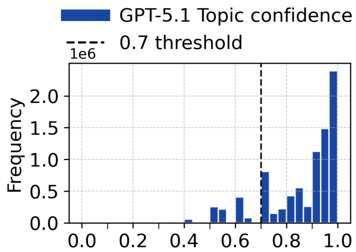
Panel A: Post Volume by Sentiment and Labeling Approach



Panel B: Distribution of *sentiment confidence*



Panel C: Distribution of *topic confidence*



Note: Panel A shows the evolution of daily post counts by sentiment label across the three classification approaches: FinBERT, ChatGPT-4o and GPT-5.1. Positive, neutral and negative posts are shown as stacked daily counts. FinBERT assigns a relatively larger share of posts to the neutral category, whereas the ChatGPT-based classifications identify a higher share of positive and negative sentiment. While the ChatGPT-4o classification is based on post titles only, the GPT-5.1 classification uses both post titles and selftext, i.e. the body of the post, providing additional contextual information for the classification task. FinBERT and ChatGPT-4o cases include only the eight investing and crypto subreddits, whereas in the GPT-5.1 case all subreddits listed in Table 1 are included. Panel B reports the distribution of sentiment confidence scores for ChatGPT-4o and GPT-5.1, while Panel C reports the distribution of GPT-5.1 topic confidence scores; topic confidence scores were not produced for the ChatGPT-4o classification. Confidence scores range from 0 to 1 and reflect the model's internal certainty in the assigned label. The dashed vertical line marks the 0.7 confidence threshold. Post counts are in millions.

ChatGPT-4o was applied to post titles only, whereas GPT-5.1 used titles and post body text. The stronger dispersion of GPT-5.1 classifications across positive and negative labels therefore suggests that additional context enables a more differentiated classification, rather than proving that GPT-5.1 is strictly superior. At the same time, ChatGPT-4o's sentiment-confidence distribution appears somewhat less dispersed, plausibly because the title-only setup produces more neutral classifications, for which the model may be relatively more certain. No direct confidence comparison with FinBERT is possible, as FinBERT does not provide an equivalent certainty score or sentiment-strength measure. For the GPT-based classifications, most sentiment-confidence scores exceed the 0.7 threshold. For GPT-5.1, topic-confidence scores are somewhat more concentrated at high values than sentiment-confidence scores, suggesting that topic classification is easier than sentiment classification. This is intuitive, as sentiment is often more nuanced, ambiguous and context-dependent. While these confidence values cannot be externally validated, they indicate high internal model certainty and suggest that filtering mainly removes lower-confidence borderline cases rather than a large share of the sample.

To further assess labelling quality, we conducted both qualitative and quantitative checks. Qualitatively, we manually examined a small subset of posts and asked the LLM to provide its reasoning for the assigned sentiment or topic label. The model's explanations appeared coherent and context-aware, also in cases involving humor, informal language or implicit market views. Quantitatively, we conducted a repeated-labelling exercise to assess the internal consistency of the final GPT-5.1 classification (see Appendix E). We selected all December 2018 and December 2025 posts across all included subreddits (73,669 posts). These two months were chosen to include one pre-COVID and one post-COVID window; the exact selection was otherwise arbitrary and should not mechanically affect the consistency assessment. Using the same prompt and model settings, GPT-5.1 independently classified the same post sample ten times. Each repeated classification was then compared with the baseline classification. The repeated-labelling exercise indicates high internal consistency for the main categorical variables. Across the full sample, the sentiment label matched the baseline in 88.9% of baseline-versus-repeat comparisons, while the topic label matched in 91.1%. At the row level, 79.3% of posts received the same sentiment label in all ten repeated runs, and 84.7% received the same topic label in all ten runs. Ticker classifications were similarly stable, matching the baseline in 91.4% of comparisons, while region classifications matched in 87.7%. Importantly, sentiment disagreements are mainly concentrated around the neutral category rather than direct polarity reversals. Across all baseline-versus-repeat comparisons, direct positive-negative contradictions account for only around 0.5% of comparisons, while most deviations involve transitions between a directional label and neutral. At the row level, only 807 posts, or 1.1% of the sample, exhibited at least one direct positive-negative contradiction across the ten repeated runs. Neutral-related shifts were more common, affecting 14,787 posts, or 20.1% of the sample, at least once. This suggests that remaining instability mainly reflects borderline cases around weak or ambiguous sentiment, rather than systematic disagreement about sentiment direction. Stability was lower for more interpretative auxiliary fields, such as trade direction, market emotion and time horizon. This is expected, as these variables are harder to infer from Reddit posts and often require contextual interpretation beyond the literal text. The lower consistency of these auxiliary variables therefore reflects the more demanding nature of the task, rather than instability in the core sentiment and topic classifications. Applying a confidence-based filter further improves the stability of the main sentiment and topic classifications. Among posts with sentiment and topic confidence of at least 0.7 in each case, sentiment-label agreement rises to 89.8%, while topic-label agreement increases to 95.6%. The improvement is particularly pronounced for topic labels, suggesting that the model's confidence score helps identify more stable classifications. For the auxiliary fields, however, the gains are more limited and not uniform: some variables improve modestly, while others show little change or even slightly lower agreement after filtering. This reinforces the interpretation that these fields are intrinsically harder to infer from Reddit posts and should be used with caution, if at all, especially in quantitative analyses. Given the scale of the data and the risk of noisy borderline labels, we therefore use confidence-based restrictions as a conservative robustness filter for the core sentiment and topic variables. This allows us to assess whether the main results are preserved when focusing on observations for which the model reports a minimum degree of certainty.

In addition, posts classified under the topic category *other* were excluded to ensure that only investing-related and thematically relevant discussions are captured in the R-RISI. This category typically contained off-topic content, such as memes, jokes, personal remarks, or general lifestyle discussions, which are unlikely to reflect genuine investment sentiment. Removing these posts helps reduce noise, enhances the economic interpretability of the indicator, and ensures that the resulting sentiment measure represents discussions directly linked to market activity and trading behaviour. Topic labeling was applied strictly according to the predefined set of categories.

Overall, the GPT-based labeling approach proved to be a powerful tool for handling and "understanding" linguistic nuance, sarcasm and non-literal expressions, allowing for a more accurate classification of sentiment and topics in Reddit posts than FinBERT. FinBERT was therefore not further used for the construction of the indicator, given its limited ability to capture contextual meaning and its tendency to assign neutral labels when textual cues are ambiguous. The GPT-5.1-labeled dataset was used instead of the GPT-4o-labeled dataset, as it incorporates additional textual context from post body text. The final dataset therefore consists of high-confidence sentiment and topic classifications for 4,156,692 posts[1] for the selected subreddits (see Table 1) from January 2015 to April 2026.

4.2 Construction of the Reddit Retail Investor Sentiment Indicator (R-RISI)

The construction of R-RISI requires transforming individually labeled Reddit posts into a consistent daily indicator that captures the prevailing tone of retail investors' discussions. The aim is to quantify fluctuations in optimism and pessimism expressed on Reddit, providing a proxy for shifts in retail investor sentiment.

A simple way to aggregate social-media sentiment would be to count the daily number of positive and negative posts and take their difference. While easy to compute, this approach is highly sensitive to changes in overall posting activity. For instance, spikes in the total number of posts during periods of market stress can distort the signal, even if the underlying sentiment balance remains unchanged. Moreover, relying solely on raw counts treats all posts as equally important, ignoring that some messages attract vastly higher attention, visibility, and engagement than others.

A first improvement is to normalise the counts by total post volume – either including or excluding neutral posts – so that the measure reflects the share of positive and negative tones rather than absolute levels. However, including neutral posts in the normalisation can bias the analysis when the share of neutrally labeled posts fluctuates sharply, for example when many informational or unrelated posts are published at times.

A superior approach is to use z-score standardisation, which expresses each day's activity relative to its historical mean and dispersion. The z-score shows how unusual a day's positive or negative engagement is compared with the subreddit's time history, accounting for structural differences in posting frequency. The 365 day window is chosen to balance noise reduction and timeliness, as it smooths short lived bursts in activity, spans full annual seasonality in retail attention and market calendars, and still allows the indicator to adapt to persistent regime shifts without being dominated by a small number of extreme episodes. To further reflect that not all posts contribute equally to investor sentiment, our R-RISI applies engagement-based weighting,

[1] The total count of posts can be split in the following counts per subreddit — r/wallstreetbets: 1,288,173; r/CryptoCurrency: 1,062,675; r/Bitcoin: 575,938; r/CryptoMarkets: 197,584; r/stocks: 157,266 r/ethereum: 135,032; r/investing: 107,921; r/StockMarket: 80,059; r/Economics: 97,422; r/geopolitics: 41,731; r/energy: 50,518; r/Gold: 40,481; r/Wallstreetsilver: 278062; r/options: 43,830.

assigning higher influence to posts that triggered more attention and discussion. This double standardisation – first across engagement intensity, then across time – enables the R-RISI to isolate sentiment shifts that are both economically and statistically meaningful. As a result, the indicator provides a robust, scale independent measure of changes in retail investor mood. This allows a direct comparison across subreddits and time as well as to standardized market data.

The indicator construction follows three main steps: (i) aggregating posts into daily sentiment totals and weighting posts by engagement intensity, and (ii) standardising the results through rolling z-scores and (iii) calculating net sentiment.

(i) Aggregating post-level sentiment and weighting by engagement

Each Reddit post i has been classified into one of three sentiment labels $L_i \in \{positive, negative, neutral\}$ using OpenAI's large language model ChatGPT-4o and 5.1 (see Section 4.1). Posts are then grouped by day t and subreddit s, producing three daily totals per subreddit, i.e. one per sentiment label. In the unweighted case, every post contributes equally:

$$N_{t,s,l} = \sum_{i \in (t,s,l)} 1, \tag{1}$$

where l denotes the sentiment label. This yields the raw number of positive, negative, and neutral posts per day.

While post counts capture the overall direction of sentiment, they ignore how strongly posts resonated within the community. In practice, a small number of highly visible posts can attract disproportionate attention and shape the prevailing tone of discussions. To capture this, each post i is assigned an engagement weight $w_i^{(k)}$, where k denotes the weighting scheme applied.

With the available Reddit data, two engagement variables are observable – the variable *score* and the number of comments *num_comments* a post receives. Based on these, four alternative weighting schemes are defined:

$$\begin{gathered} w_i^{(\text{score})} = \text{score}_i, \quad w_i^{(\text{comments})} = \text{comments}_i, \quad w_i^{(\text{total interaction})} = \text{score}_i + \text{comments}_i, \\ w_i^{(\text{react})} = \ln(1 + \text{score}_i) + \ln(1 + \text{comments}_i). \end{gathered} \tag{2}$$

These weights reflect different aspects of community engagement:

- *score*: the number of upvotes minus downvotes proxies the level of agreement with a post,
- *comments*: the number of comments proxies the degree of discussion intensity,
- *total interaction*: the sum of score and comments captures overall visibility and attention a post received and proxies total interaction with a post in the community, and
- *react*: applies a ln(1+x) transformation to both score and comments before summing them to capture the overall engagement of Redditors with a post while mitigating extreme, abnormal reactions to few posts.

For our R-RISI, we select $w_i^{(\text{react})}$ and refer to it as reactiveness weighting, as it best approximates how reactive the retail investor crowd was to individual posts, i.e. how much collective attention a given message attracted on a particular day. Posts that generated stronger engagement (more upvotes and comments combined) are thus given proportionally higher influence in the daily sentiment computation. To control for extremely popular posts that could otherwise dominate or inflate the daily sentiment measure, we apply a ln(1+x) transformation to both score and comments before summing them. This logarithmic scaling dampens the influence of outliers – for instance, preventing a single highly engaged post from outweighing numerous moderately engaged ones and biasing the overall sentiment drift. The addition of one inside the logarithm ensures that posts with zero engagement remain defined. This adjustment is standard in the literature and minorly affects posts with very low interaction.

Hence, for each day t, subreddit s, and sentiment label l, the engagement weighted sentiment totals are computed as:

$$W_{t,s,l}^{(react)} = \sum_{i \in (t,s,l)} w_i^{(react)} \tag{3}$$

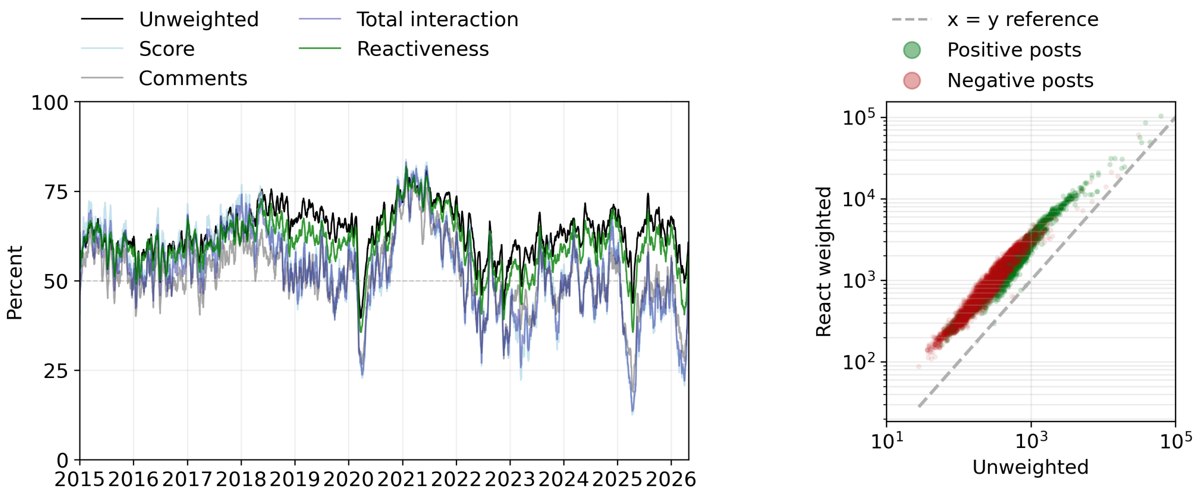
If all posts are weighted equally ($w_i^{(k)} = 1$), this collapses to the simple unweighted post count $N_{t,s,l}$. When engagement varies substantially, the weighting amplifies the tone of posts that attracted the most community attention.

However, comparing the above-defined popularity-weighting schemes with the unweighted series shows that weighting does not fundamentally alter the overall pattern of Reddit post sentiment. The measures move closely together, although engagement weighting can slightly adjust the strength or amplitude of sentiment swings at times (see Figure 6, Panel A). A comparison of daily unweighted and react-weighted sentiment counts (see Figure 6, Panel B) shows that for both positive and negative posts, all observations lie above the one-to-one reference line, indicating that engagement weighting systematically increases measured sentiment totals. The green cloud of positive posts runs roughly parallel and consistently above the reference line, implying that engagement grows proportionally with posting activity and systematically amplifies optimistic sentiment. In contrast, the red cloud of negative posts appears more concentrated and less vertically stretched, suggesting that engagement with negative content increases less strongly once overall activity rises. Overall, highly interactive posts amplify sentiment in both directions, with a somewhat stronger and more persistent effect on positive discussions, which tend to attract higher user engagement and hence exert a greater influence on the R-RISI.

Figure 6: Comparison of Popularity Weighted and Unweighted Reddit Posts

Panel A: Share of Positive Sentiment under Different Weighting Schemes

Panel B: Weighted versus Unweighted Post Counts



Note: Panel A compares the daily share of positive Reddit posts across five weighting schemes: unweighted, score-weighted, comments-weighted, weighted by their sum (total interaction) and react-weighted (reactiveness). The data are shown as 14-day rolling averages to smooth short-term noise. Panel B contrasts unweighted and react-weighted totals for positive and negative posts on a logarithmic scale. Each dot represents one day; the dashed line indicates equality between weighted and unweighted totals.

(iii) Rolling z-score standardisation

To control for long term shifts in posting intensity and engagement levels, each weighted sentiment series is transformed into a 365-day z-score within each subreddit or aggregated for a specific subreddit or list of subreddits:

$$z_{t,s,l}^{(\text{react})} = \frac{W_{t,s,l}^{(\text{react})} - \mu_{t,s,l,365}^{(\text{react})}}{\sigma_{t,s,l,365}^{(\text{react})}}, \tag{4}$$

where $\mu^{(\text{react})}_{t,s,l,365}$ and $\sigma^{(\text{react})}_{t,s,l,365}$ denote the rolling mean and standard deviation of the weighted daily totals over the previous 365 days. This expresses each day's sentiment in standard deviation units relative to its historical baseline, ensuring comparability across time.

(ii) Net sentiment calculation

Finally, the net sentiment for each subreddit and day is defined as the difference between the z-scored positive and negative sentiment components:

$$\text{R-RISI}^{(\text{react})}_{t,s} = z^{(\text{react})}_{t,s,\text{positive}} - z^{(\text{react})}_{t,s,\text{negative}}. \tag{5}$$

A positive value on a given day indicates that positive posts received above-average engagement relative to their historical norm compared to negative posts, signalling an unusually optimistic tone. Conversely, negative values reflect periods of heightened engagement with bearish content. This design ensures that the R-RISI captures relative sentiment shifts rather than absolute activity changes. Weighting by engagement intensity highlights the most influential posts shaping discussions, while z-scoring adjusts for structural changes over time.

To enable a direct comparison with financial market developments, z-scores are also computed for key market variables such as equity indices, individual equity or asset prices.

The resulting daily z-score series is naturally quite volatile, as retail discussions on Reddit can fluctuate sharply in response to news events or social media dynamics. 30-day rolling averages of the z-scores of both sentiment and market variables improve qualitative interpretability. The smoothing window, however, depends on the time horizon one wants to visually look at our R-RISI. Going several years into the past, an unsmoothed version is not sensible as the daily volatility is too high. Looking at a recent period of few years or some months then e.g. a seven day smoothing may fit better. We align the standardised R-RISI with market data to assess whether shifts in online tone coincide with periods of market exuberance, stress, or correction.

5 Results

In this section, we discuss the results from the LLM based labelling and describe the composition of Reddit discussions across major investing- and crypto-related subreddits. We analyse the resulting R-RISI measure over time for different samples of underlying equities and compare the obtained retail investor sentiment with market prices and alternative retail sentiment measures.

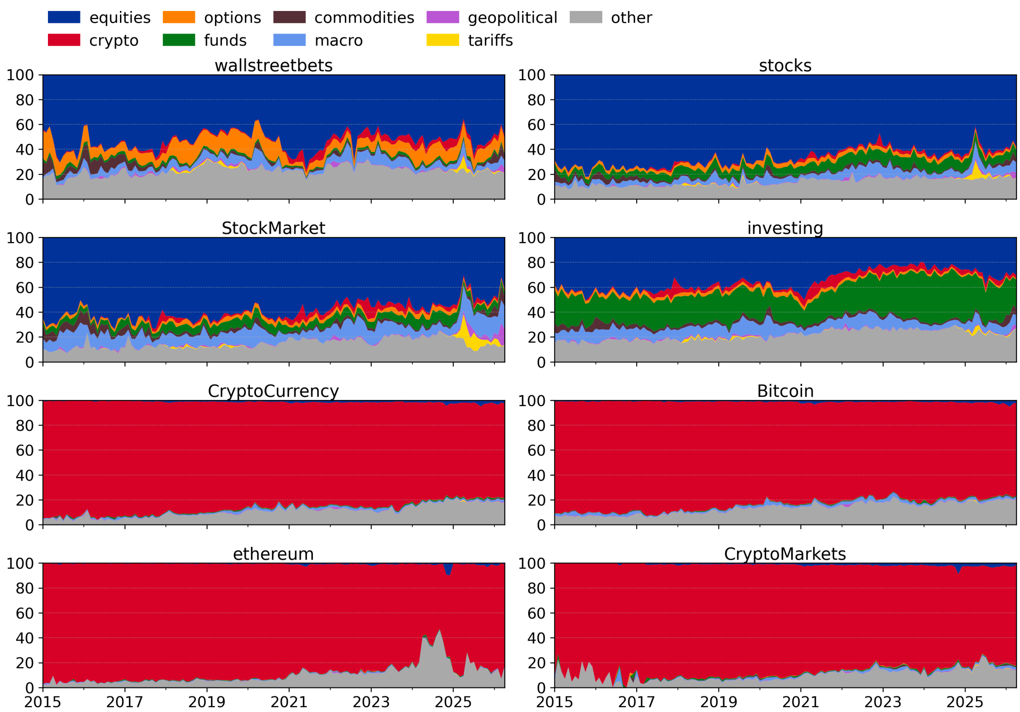
5.1 Descriptive Results from Reddit Discussions

Figure 7 shows the thematic composition of discussions across the eight major investing- and crypto-related subreddits included in our sample between January 2015 and April 2026. The topic composition for the remaining subreddits is in Appendix F. Equity-related topics account for the largest share of posts within the investing-oriented subreddits throughout the full sample period, although the relative importance of other themes differs across communities. r/wallstreetbets displays a broad topic composition, with roughly half of the discussions concerning equities and a persistent presence of posts related to options and macroeconomic developments, alongside a relatively stable share of around 20 percent of posts classified as *other*—the highest share across all subreddits. r/investing shows a steady dominance of fund-related discussions, which from 2022 onwards represent the largest topic category of posts. r/StockMarket features a more balanced distribution across topics, though equities remain the leading theme, followed by macroeconomic discussions, a pattern similarly observed for r/stocks, but here fund discussions come second. Within the investing-related subreddits, a small but visible share of posts referring to crypto assets emerges from around 2020 onward. In contrast, the crypto-related subreddits are almost exclusively devoted to cryptocurrency topics, with equity-related content accounting for only a negligible fraction that becomes slightly more visible in later years. Subreddits such as r/Bitcoin and r/ethereum follow comparatively strict moderation rules, which contribute to maintaining a narrow thematic focus. The share of *other* discussions within these communities ranges between ten and 30 percent and shows a gradual increase over time.

A particularly notable topical focus emerges around April 2025, when the share of posts related to tariffs rises sharply across several investing-oriented subreddits, coinciding with trade-policy tensions and tariff measures introduced by the Trump administration. This contemporaneous adjustment in discussion topics illustrates the responsiveness of retail investor communities to unfolding market events. The magnitude and timing of this reaction suggest that Reddit discussions are highly topical and adapt rapidly to new information, providing further indication that online retail investor discourse closely mirrors prevailing market narratives.

Reddit discussions on financial assets exhibit a clear concentration around a limited number of prominent tickers across both equity and crypto subreddits. Figure 8 illustrates the most frequently mentioned tickers in investing (Panel A) and crypto-related (Panel B) communities. In investing subreddits, discussions are dominated by a handful of equities well-known to be discussed among retail investors, including Tesla (TSLA), Gamestop (GME), Nokia (NOK), and Apple (AAPL), Palantir (PLTR) alongside technology leaders such as Nvidia (NVDA) and Meta/Facebook (META) and broader indices like S&P 500 ETF (SPY) or SPX. This concentration suggests that retail investors' attention tends to cluster around a few highly visible stocks. In crypto subreddits, conversations are centered almost exclusively on Bitcoin (BTC) and Ethereum (ETH), with other tokens such as DOGE, Ripple (XRP), and Solana (SOL) featuring more sporadically. Mentions of stablecoins (USDT, USDC) appear less frequently and, when they occur, predominantly refer to U.S.-dollar- denominated tokens. The two panels highlight how both equity and crypto discussions are driven by a small set of recurring names, reflecting a collective focus on well-known and sentiment-sensitive assets.

Figure 7: Reddit Discussion Topics Across Key Investing and Crypto Subreddits Over Time



Note: The chart depicts the share of Reddit posts per topic category, expressed as a percentage of all posts within each subreddit. Post counts were aggregated at a monthly frequency. Topic categories were assigned using OpenAI's GPT-5.1 model, as in Section 4.1. The data cover the period from January 2015 to April 2026. See remaining subreddits in Appendix F.

Taken together, the topic composition and ticker-level analysis illustrate the main themes and assets that dominate retail investor discussions on Reddit. Building on these insights, the following section introduces the R-RISI, which quantifies the tone of these discussions over time.

Figure 8: Most Frequently Mentioned Tickers Across Investing and Crypto-related Subreddits

Panel A: Investing subreddits

Panel B: Crypto subreddits

Note: Ticker mentions were extracted using ChatGPT-5.1, which relied on context to identify them regardless of whether they were mentioned as standard tickers, abbreviations, full names, or misspelled text in the Reddit posts. Word frequencies were aggregated over the full sample period (January 2015 to April 2026). Panel A covers all four investing subreddits. Panel B includes only the broader crypto subreddits r/CryptoCurrency and r/CryptoMarkets, excluding r/Bitcoin and r/ethereum to prevent extreme domination. Limit to 200 words each.

5.2 Sentiment Indicator

A first step is to understand the extent to which sentiment dynamics differ across online communities. The correlation matrix in Figure 9 Panel A reveals clear segmentation across subreddits. Sentiment in investing-oriented communities (r/wallstreetbets, r/investing, r/stocks, r/StockMarket) shows moderate positive co-movement, consistent with a shared focus on equity markets and macro-financial news. Within the crypto segment, sentiment is strongly correlated between broad-interest subreddits (r/CryptoCurrency and r/CryptoMarkets) but largely uncorrelated for asset-specific discussions on r/Bitcoin and r/Ethereum. Cross-group correlations between finance and crypto communities remain close to zero, indicating that these groups respond to distinct information sets and that sentiment formation on Reddit is segmented along asset-class lines. The generally positive but moderate correlations indicate that subreddit-level sentiment measures share a common retail-investor component while preserving meaningful cross-community heterogeneity.

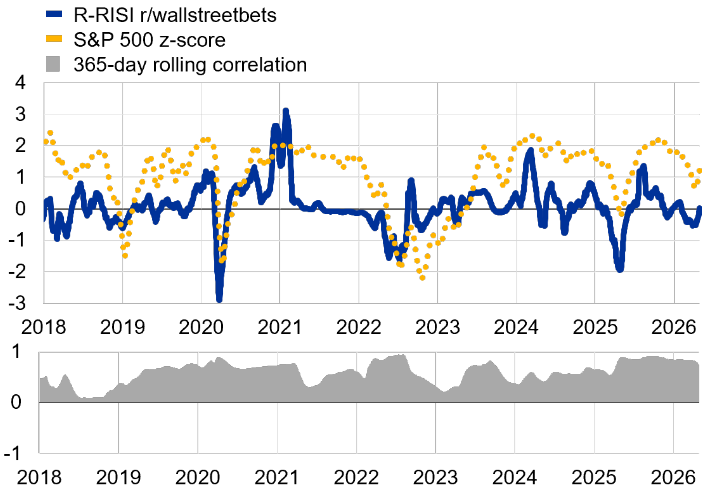
To evaluate whether sentiment extracted from Reddit meaningfully reflects broader market conditions, Figure 9 Panel B compares net sentiment on r/wallstreetbets ($R\text{-}RISI_{SPX}$) with S&P 500 equity market dynamics. The two series move closely together, with sentiment increasing

Figure 9: R-RISI versus Market Developments and Alternative Retail Investor Indices

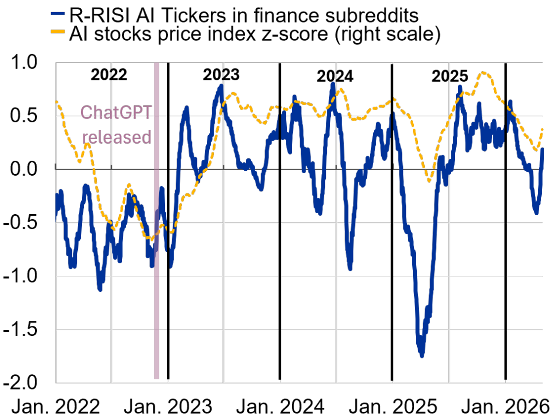
Panel A: R-RISI Subreddit Correlation

	wallstreetbets	stocks	StockMarket	investing	CryptoCurrency	Bitcoin	Ethereum	CryptoMarkets
wallstreetbets	1.00							
stocks	0.40	1.00						
StockMarket	0.36	0.36	1.00					
investing	0.36	0.29	0.31	1.00				
CryptoCurrency	0.14	0.13	0.16	0.13	1.00			
Bitcoin	0.15	0.06	0.09	0.09	0.36	1.00		
Ethereum	0.01	0.00	0.04	0.03	0.16	0.08	1.00	
CryptoMarkets	0.22	0.19	0.22	0.20	0.37	0.26	0.13	1.00

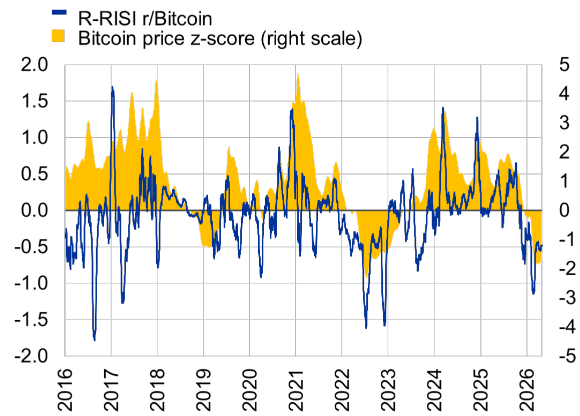
Panel B: $R\text{-}RISI_{SPX}$ and its Correlation with the S&P 500



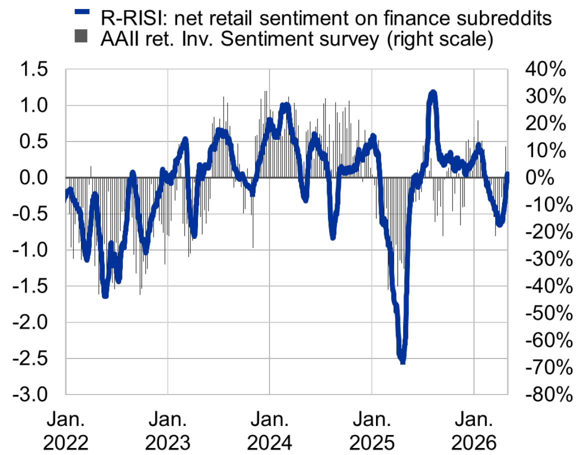
Panel C: $R\text{-}RISI_{AI}$ versus custom AI index



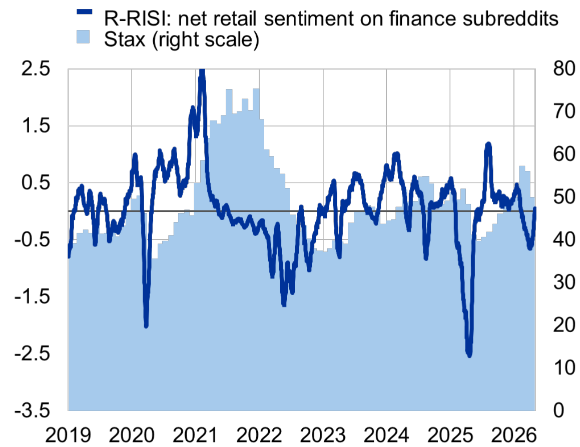
Panel D: $R\text{-}RISI_{BTC}$ versus Bitcoin



Panel E: R-RISI versus AAII survey



Panel F: R-RISI versus STAX



Note: Panel A shows the correlation of reactiveness-weighted net sentiment across subreddits. Panel B: net retail sentiment is from r/wallstreetbets posts. Posts are GPT-5.1 topic-classified (equities, funds, macro, options, commodities, geopolitics, tariffs, crypto), non-investing content removed, and labelled positive, negative or neutral. Net sentiment is the difference between 365-day z-scores of positive and negative posts, weighted by log-comments and log-upvotes. The S&P 500 shows the 365-day z-score of index levels; both series are 30-day smoothed. The lower subpanel reports 365-day rolling correlation. Panel C uses all investing subreddits. The AI index is a market-cap-weighted basket of 41 S&P 500 AI tickers (Bloomberg CIX); Posts are filtered for these tickers (list in Appendix D). Panel D: R-RISI uses all posts in r/Bitcoin. Panel E: R-RISI aggregates r/wallstreetbets, r/investing, r/stocks and r/StockMarket. The AAII Survey is the weekly spread of bullish minus bearish retail views for the next six months. Panel F: R-RISI as in E. STAX is the Charles Schwab Trading Activity Index.

during market upswings and declining during periods of stress. Rolling 365-day correlations rise substantially during episodes of heightened financial uncertainty and recede in more tranquil periods. This pattern is consistent with episodic procyclicality among speculative retail investors, who appear to anchor more strongly to prevailing price dynamics when market volatility is high[2].

The flexibility of the R-RISI framework allows sentiment to be isolated for specific assets or topic categories. Figure 9 Panel C focuses on firms associated with artificial intelligence, filtering Reddit posts for references to 41 AI-related S&P 500 constituents (see Appendix D). Sentiment for this subset ($R\text{-}RISI_{AI}$) closely tracks the performance of a custom market-capitalisation-weighted AI stock index. Short-term fluctuations are more pronounced than in the price series, yet the indicator captures key technology-related events, including a marked positive spike coinciding with the public release of ChatGPT.

Crypto-specific sentiment is more volatile. Figure 9 Panel D compares BTC sentiment in crypto subreddits ($R\text{-}RISI_{BTC}$) with Bitcoin price z-scores. The sentiment series displays frequent and sizable swings, reflecting the underlying speculative nature of the asset. Importantly, $R\text{-}RISI_{BTC}$ captures prolonged negative sentiment during major crypto crash periods in 2017-18 and 2022, demonstrating its ability to mirror broad market conditions even in highly volatile environments. We present R-RISI for all selected subreddits without smoothing and with 7-day and 30-day smoothing in Appendix G.

5.3 Benchmarking R-RISI against established retail sentiment measures

To clarify how R-RISI should be interpreted relative to established retail sentiment benchmarks, we situate the Reddit cohort underlying R-RISI against the reference populations of the AAII Survey and the Charles Schwab Trading Activity Index (STAX) described in Section 3.2.

Reddit's investing and crypto communities skew substantially younger, more male and more speculative than the broader retail investor population. According to Aubin and Liedke (2025) and Hernandez (2025), approximately 65 percent of U.S. Reddit users are aged between 18 and 39, about 63 percent are male and the median household income is close to USD 75,000. By comparison, AAII survey respondents are on average above 60 years old with multi-decade investing experience, and Charles Schwab retail clients span a wider wealth and age range while

[2] A slight variation of this figure has also been published by us in Box 2 of the November 2025 European Central Bank Financial Stability Review, comparing sentiment in r/wallstreetbets to global equity markets using the MSCI World Index. The publication is accessible at https://www.ecb.europa.eu/press/financial-stability-publications/fsr/focus/2025/html/ecb.fsrbox202511_02~3fb190ec50.en.html.

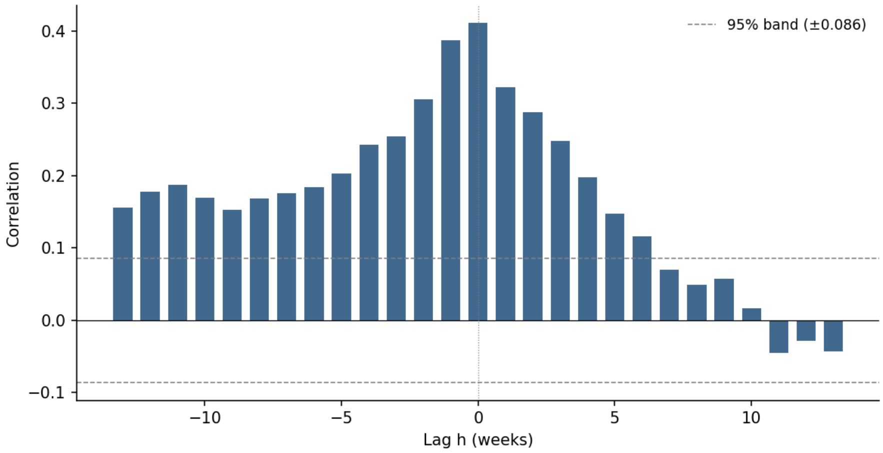
tending toward long-horizon asset allocation (Reichenbach & Walther, 2023). R-RISI should therefore be understood as a measure of the speculative, online-active cohort of retail investors rather than of the retail population as a whole. We view this as a deliberate design feature: the cohort captured by R-RISI is precisely the segment whose attention and trading behaviour have most visibly shaped market dynamics over the sample period, from the 2020–21 meme-stock episode through the 2022 crypto drawdown to the 2025 tariff correction.

Figure 9 Panels E and F provide a visual comparison of R-RISI with the AAII Sentiment Survey and the Charles Schwab Trading Activity Index (STAX). In Panel E, R-RISI and the AAII bull-bear spread both exhibit broadly similar cyclical patterns, although the R-RISI responds more rapidly to market events likely due to its daily frequency. This suggests that R-RISI may provide a more timely signal of shifts in retail investor mood. Panel F benchmarks the R-RISI against the STAX, displaying broadly similar directional patterns, albeit mostly around major market events. The divergence around the year 2021, where STAX rises while Reddit sentiment turns more cautious, may reflect differences in user composition: Schwab clients are wealthier and more long-term oriented, whereas Reddit users appear more short-term and speculative.

To investigate the timing relationship between R-RISI and the AAII Bull–Bear spread, we aggregate the S&P-oriented daily R-RISI to a weekly frequency aligned with the AAII publication schedule, over the window January 2016 to December 2025, resulting in 520 weekly observations. The cross-correlation function $\rho(h)$ displayed in Figure 10 shows the correlation between R-RISI at week $t + h$ with the AAII at week t. AAII is strongly correlated with past realisations of R-RISI and the strength of the relationship is increasing with h. The correlation of R-RISI with future values of AAII is weaker and declines swiftly. The asymmetry between the strength of the correlation of one week in the past (0.39) and one week in the future (0.32) indicates that R-RISI leads AAII by approximately one week. The asymmetry is consistent with Reddit discussions responding to market developments slightly before AAII respondents adjust their stated views, as one would expect from the higher posting frequency and the more reactive nature of online discussions. Taken together, the lead–lag results suggest that R-RISI carries information about the direction of AAII sentiment that is not yet impounded in the survey itself, at least at weekly horizons.

As an out-of-sample test of the informational content of R-RISI, we nowcast the AAII Bull–Bear spread for the following week using two autoregressive specifications: a baseline AR(1) model using only on the lagged AAII alone, and an augmented AR(1) model that additionally

Figure 10: Cross-Correlation Function: R-RISI and AAII at Weekly Lags



Note: Cross-correlation function $\rho(h) = \text{corr}(\text{R-RISI}_{t+h}, \text{AAII}_t)$ at weekly frequency. Dashed lines denote approximate 5% significance bands.

includes R-RISI. The sample is split 82/18 into a pre-2024 training window and a hold-out test window; z-scores are computed using full-sample moments. Adding R-RISI reduces the out-of-sample root mean squared error (RMSE) from 0.1025 to 0.1013 and the mean absolute error (MAE) from 0.0816 to 0.0797. The coefficient on R-RISI in the augmented model is 0.0116, providing mild evidence that R-RISI contains information about near-term AAII sentiment beyond what AAII's own autoregressive structure captures.

Table 4: Out-of-sample nowcast of next-week AAII, with and without R-RISI

Model	R-RISI Coefficient	RMSE	MAE
AR(1) baseline		0.1025	0.0816
AR(1) + R-RISI	0.0116	0.1013	0.0797

Note: The table reports statistics on the out-of-sample nowcasting exercise to establish the information content in today's R-RISI that influences next-week AAII. The first row indicates the forecasting ability of a simple AR(1) model, while the second row extends that model by R-RISI. RMSE = Root Mean Squared Error. MAE = Mean Absolute Error.

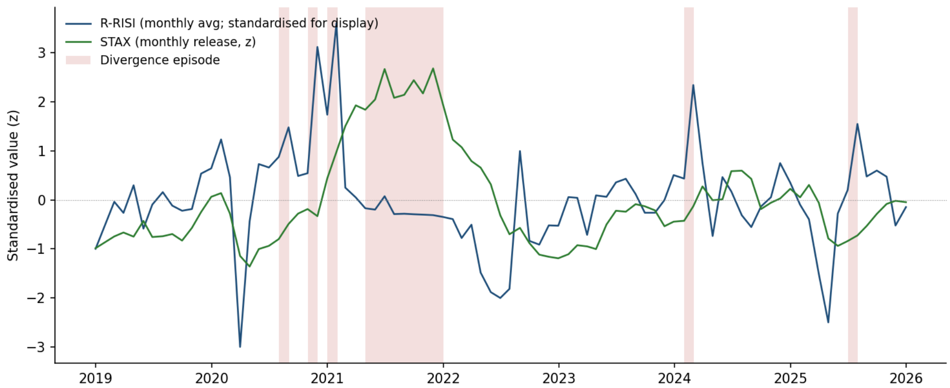
Unlike AAII, which captures stated views from a self-selected panel of long-horizon individual investors, STAX reflects realised trading behaviour of Charles Schwab retail clients whose wealth and age distributions differ substantially from the Reddit cohort underlying R-RISI. In line with the visual evidence from Figure 9 Panel E, we therefore expect R-RISI and STAX to broadly co-move but to diverge episodically whenever the speculative online cohort takes positions at odds with the broader Schwab retail base. To identify such episodes, we aggregate R-RISI to a monthly frequency, align it with the monthly STAX release, and compute the spread between their z-scored values over 84 monthly observations (January 2019 to December 2025).

Thirteen months out of 84 exceed a $1.5 \cdot \sigma$ threshold on the absolute spread (Figure 11). These episodes cluster around three interpretable regimes: the November 2020 to January 2021 meme-stock rally, in which, although both indicators move in the same direction at the same time, R-RISI turned markedly more positive than STAX, which reflected an overall market recovery; a period from mid-2021 into early 2022 in which retail brokerage rotation into stable holdings drove STAX upward while Reddit sentiment stayed cautious; and a brief episode in 2025 associated with tariff-policy uncertainty and mid-year reflation discussions. The episode list is robust to the threshold choice: 26 episodes at σ and four at $2 \cdot \sigma$ preserve the same three-regime clustering that may span wider and narrower periods respectively.

These episodes validate the interpretation of R-RISI as a measure of the speculative online cohort: the indicator moves sharply when that cohort reacts to news or social dynamics in ways that the broader Schwab retail base does not, and the few episodes it isolates are those previously identified as retail-driven.

The results together suggest that R-RISI, AAII, and STAX are complementary indicators rather than substitutes. Table 5 compares the three measures along key dimensions. R-RISI captures what speculative-retail participants are *saying* on a continuous, daily basis; AAII captures what a self-selected panel of long-horizon investors *expects* over the next six months; and STAX captures what the broad Schwab retail book is *doing* via realised portfolio positioning.

Figure 11: Divergence Episodes between R-RISI and STAX, Monthly, 2019–2025



Note: Monthly z-scored R-RISI and STAX over 2019–2025. Shaded bands mark months where the absolute spread exceeds $1.5 \cdot \sigma$ (13 of 84 months). Divergence episodes cluster around: (i) November 2020 – January 2021 meme-stock rally; (ii) mid-2021 – early 2022 retail rotation; (iii) the 2025 tariff correction and reflation discussion.

Table 5: Positioning Matrix: R-RISI, AAII Sentiment Survey, and STAX

	R-RISI (Reddit)	AAII Survey	STAX (Schwab)
Nature	LLM-based speech measure	Survey-stated expectations	Realised trading behaviour
Frequency	Daily	Weekly	Monthly
Population	Speculative online retail	Self-selected long-horizon retail	Broad Schwab brokerage clients
Median age	23–35 years	50–60+ years	45–65 years
Investment horizon	Short-term / speculative	Six-month horizon	Long-term positioning
Role in paper	High-frequency signal	Lead–lag benchmark	Divergence benchmark

Note: Comparison of the three retail sentiment indicators. The AAII Sentiment Survey is published weekly by the American Association of Individual Investors as the net-bullish (bullish% − bearish%) spread. STAX is the Charles Schwab monthly Trading Activity Index reflecting realised client portfolio positioning. Demographic estimates from Aubin and Liedke (2025), Hernandez (2025), and Reichenbach and Walther (2023).

6 Empirical analysis

Although R-RISI and equity prices are strongly correlated, it is unclear whether the extracted sentiment reflects actual retail trading. In the absence of detailed data on retail trading activity, we test the effect of R-RISI on prices, interpreted as trading outcomes. For robustness, we also test the relationship between sentiment and trading volumes on the S&P 500 (see Appendix H).

6.1 Testing relationship between sentiment and prices

Reddit's granularity allows analysis across different aggregation levels in terms of topics, investment specifications and asset classes. Our strategy involved moving from a more general assessment of retail sentiment, i.e. retail sentiment about the overall market against a market index, to more specific cases, like a selection of equities of a single price.

We begin with broad retail sentiment from r/wallstreetbets. As r/wallstreetbets often discusses large-cap U.S. stocks, this captures general investor mood relevant to the S&P 500 (SPX) movements. This setup serves as our benchmark case for broad market sentiment, as the market value weighted price index itself is driven by a selection of equity prices of very large corporations. Hence our first regression will test the effect of R-RISI on the S&P 500 ($R - RISI_{SPX} \rightarrow SPX$). Second, we assess topic-specific aggregation by constructing $R - RISI_{AI}$ from posts on 41 AI stocks across four investing subreddits and match it to a custom AI index. This tests whether R-RISI reflects topical trends and their market impact within a defined thematic segment ($R - RISI_{AI} \rightarrow AI$). Finally, to show robustness across asset classes, we relate $R - RISI_{BTC}$ constructed from posts from subreddit r/Bitcoin to Bitcoin prices ($R - RISI_{BTC} \rightarrow BTC$). $R - RISI_{BTC}$ focuses on a single-asset discussed at high frequency, strengthening the link between Reddit sentiment and prices. Bitcoin also plays an increasingly important role in retail portfolios, with retail investors forming a large share of its investor base.

Table 6 reports summary statistics for the three sentiment–price samples in the upper panel. The $R\text{-}RISI_{SPX}$ and the $R\text{-}RISI_{BTC}$ samples each contain 2,307 observations, covering the period from January 2016 to April 2026. The $R\text{-}RISI_{AI}$ sample is restricted to the period from January 2023 to April 2026 to avoid a structural break caused by the popularity of AI-related discussions before and after the introduction of ChatGPT to the market in 2023. The average $R\text{-}RISI_{SPX}$ is 0.126 with an interquartile range of 0.986, close to its standard deviation of 1.211. The mean $R\text{-}RISI_{AI}$ is more neutral at 0.096 but has a slightly wider interquartile range of 0.998, respectively. The Bitcoin sample has a marginally narrower interquartile range of 0.968, which is more skewed to the negative side. Together with a mean $R\text{-}RISI_{BTC}$ of -0.049, the sample statistic suggests overall more negative sentiment on Bitcoin than on stock prices.

Our regression specification follows the strategy outlined in Tetlock (2007) and García (2013):

$$z_t^{\mathrm{p}} = \alpha + \sum_{k=1}^{5} \beta_k z_{t-k}^{\mathrm{p}} + \sum_{k=1}^{5} \gamma_k R\text{-}RISI_{t-k}^{p} + \sum_{k=1}^{5} \delta_k \mathrm{VIX}_{t-k} + \phi' \mathrm{DOW}_t + \varepsilon_t, \tag{6}$$

where z_t^p is the z-score at time t of the respective equity price index p, with $p \in \{SPX, AI, BTC\}$. $RISI_{t-k}^p$ is the R-RISI for retail sentiment at lag k in each of the three cases. Definitions follow section 3 and 4. k denotes lags; we use 5, corresponding to one trading week. VIX is a control for volatility of prices, and DOW is a vector of dummies to control for day of the week, Monday excluded. We estimate results using Newey-West HAC estmators.

Regression results for Equation 6 are shown in Table 7 Panel A. The first R-RISI lag has a positive effect on equity price z-scores at the 1% level for the SPX sample. The third, fourth, and fifth lag are all significant at the 5% level, but neutralize each other. Despite the first lag of R-RISI being also largest in magnitude in the AI sample, it lacks significance even at the 10% threshold. Considering the price impact of investor sentiment on Bitcoin yields positive and significant (at the 5% level) results in the first lag of R-RISI. In terms of magnitude, the effect in the SPX sample is the largest by an order of magnitude. The joint significance test of $R\text{-}RISI_{SPX}$ lags (Panel C, first row) is also significant, confirming that a full week of retail investor sentiment significantly affects prices. Our findings align with prior evidence on short-term sentiment effects.

To corroborate our results, we re-estimate the regressions using R-RISI innovations, residualised for market prices and R-RISI autoregression. We run the following regression:

$$R\text{-}RISI_t = \alpha + \beta_0 z_t^{\mathrm{p}} + \sum_{k=1}^{5} \beta_k z_{t-k}^{\mathrm{p}} + \sum_{k=1}^{5} \gamma_k R\text{-}RISI_{t-k}^{p} + u_t, \quad (7)$$

We plug residuals u_t into $R\text{-}RISI^p$ in Equation 6 to estimate the effect of innovations in retail sentiment on market prices. The R-RISI innovations aim to capture pure daily market sentiment without interference of past and present market prices and past sentiment. Table 7 Panel B shows results similar in magnitude to Panel A, the first R-RISI lag has a significant positive effect on equity indices.

Table 6: R-RISI sample statistics

	Levels					
	Obs.	Mean	Median	25%-quant.	75%-quant.	σ
$R\text{-}RISI_{SPX}$	2307	0.126	0.078	−0.366	0.620	1.211
$R\text{-}RISI_{AI}$	746	0.096	0.131	−0.331	0.667	0.984
$R\text{-}RISI_{BTC}$	2307	−0.049	−0.023	−0.541	0.427	1.067

Note: The table reports sample statistics for the R-RISI indicators used in the regressions. $R\text{-}RISI_{SPX}$ is based on all r/wallstreetbets posts from January 2016 to April 2026. $R\text{-}RISI_{AI}$ uses posts from four investing-related subreddits mentioning 41 artificial-intelligence-related SPX stocks from January 2023 to April 2026. $R\text{-}RISI_{BTC}$ is based on all r/Bitcoin posts from January 2016 to April 2026.

The statistically significant effect of lagged sentiment on current prices links sentiment to market activity and highlights the informational value of R-RISI. Reddit discussions reflect retail mood that can affect market prices.

Robustness of the main price–sentiment regression to the 2021 meme-stock episode and to the choice of rolling z-score window is documented in Appendix I. One might argue that regressing market prices on Reddit posts weighted by their user engagement introduces a hindsight bias. Users can engage with a Reddit post even days after publication, which makes it possible that the weights include information about price movements that were unknown at the time of the posting. This would artificially inflate the importance of posts that are inline with market price developments and bias the regression results towars a more significant relationship between R-RISI. For robustness, we tested the same regression specification on the unweighted version of the R-RISI. Results were robust to the findings in this section and are presented in Appendix H. We also tested the effect of R-RISI on total trading volumes, but found no effect (see Appendix H).

Table 7: Effect of R-RISI on market prices

Panel A: Levels						
	SPX		AI		Bitcoin	
	γ_k	p-value	γ_k	p-value	γ_k	p-value
$R-RISI_{t-1}$	0.0204	0.000	0.0122	0.105	0.0117	0.012
$R-RISI_{t-2}$	−0.0086	0.149	0.0001	0.981	−0.0016	0.772
$R-RISI_{t-3}$	−0.0093	0.024	−0.0004	0.947	−0.0064	0.157
$R-RISI_{t-4}$	0.0163	0.032	−0.0054	0.372	−0.0020	0.763
$R-RISI_{t-5}$	−0.0093	0.033	0.0037	0.576	−0.0027	0.584
Panel B: Innovations						
	SPX		AI		Bitcoin	
	γ_k	p-value	γ_k	p-value	γ_k	p-value
$R-RISI_{t-1}$	0.0117	0.020	0.0120	0.102	0.0109	0.023
$R-RISI_{t-2}$	0.0014	0.753	0.0029	0.654	0.0017	0.749
$R-RISI_{t-3}$	0.0078	0.059	0.0017	0.790	-0.0047	0.292
$R-RISI_{t-4}$	-0.0013	0.705	−0.0054	0.369	-0.0027	0.656
$R-RISI_{t-5}$	-0.0038	0.346	0.0051	0.454	-0.0037	0.503
Panel C: Wald Tests						
	SPX		AI		Bitcoin	
	F-stat	p-value	F-stat	p-value	F-stat	p-value
$\sum_{j=2}^{5} \beta_{1j} = 0$	21.256	0.001	4.372	0.497	9.908	0.078
$\sum_{j=2}^{5} \beta_{2j} = 0$	19.516	0.002	4.435	0.435	8.799	0.117

Note: The table reports estimated coefficients for the model in Equation 6. Panel A shows the coefficients β and p-values of the lags of R-RISI on the dependent variable, the z-score of market prices. Columns 1-2 refer to statistics for the $R\text{-}RISI_{SPX}$-SPX sample, from January 2016 to April 2026. Columns 3-4 refer to statistics for the $R\text{-}RISI_{AI}$-AI index sample, from January 2023 to April 2026. Columns 5-6 reports the results for the sample $R\text{-}RISI_{BTC}$-Bitcoin, from January 2016 to April 2026. Panel B shows the results of Equation 6 using the transformation of R-RISI from levels to innovations, controlling for auto-regressive terms and the effect of market prices on sentiment, as outlined in Equation 7. Panel C shows the F-statistics and p-values of Wald tests on joint significance of the coefficients of R-RISI lags.

6.2 Asymmetric effects across contraction and expansion regimes

The pooled regression in Equation (6) imposes a single sentiment-to-price coefficient across the full sample. García (2013) and Tetlock (2007) show with news-based sentiment, that the asset-price impact of content is roughly an order of magnitude larger in recessions than in booms. In this section we present tests whether retail investor sentiment measured by R-RISI carries the same regime asymmetry on prices of the S&P 500.

We define market downturns as SPX price declines of at least 5% from a peak. We can decompose the drawdown cycle into its constituents. Each downturn continues until prices fully recover to the previous peak. As a counterfactual, we also identify recovery episodes, where market prices increase from a trough of the previous market price drop. The recovery episode ends once prices return to the level at which the drawdown began. Finally, we consider the whole episode of market price drop and recovery. The result are three different regimes, that express market contraction, recovery and the full cycle: the peak-to-trough regime, the trough-to-recovery regime, and the full peak-to-recovery regime. Because the trough of a drawdown episode can only be identified after prices have risen above it, these are ex-post episode labels rather than pre-determined filters. Figure 12 shows the classification of the regimes over time. The colored bars, mark the time periods that are defined by the three different regimes from above in panels (a)–(c). Peak-to-trough contractions are significantly shorter than trough-to-recovery regimes. While this short time periods might challenge the result on R-RISI's properties to capture retail activity, one advantage is that those periods show stronger price volatility.

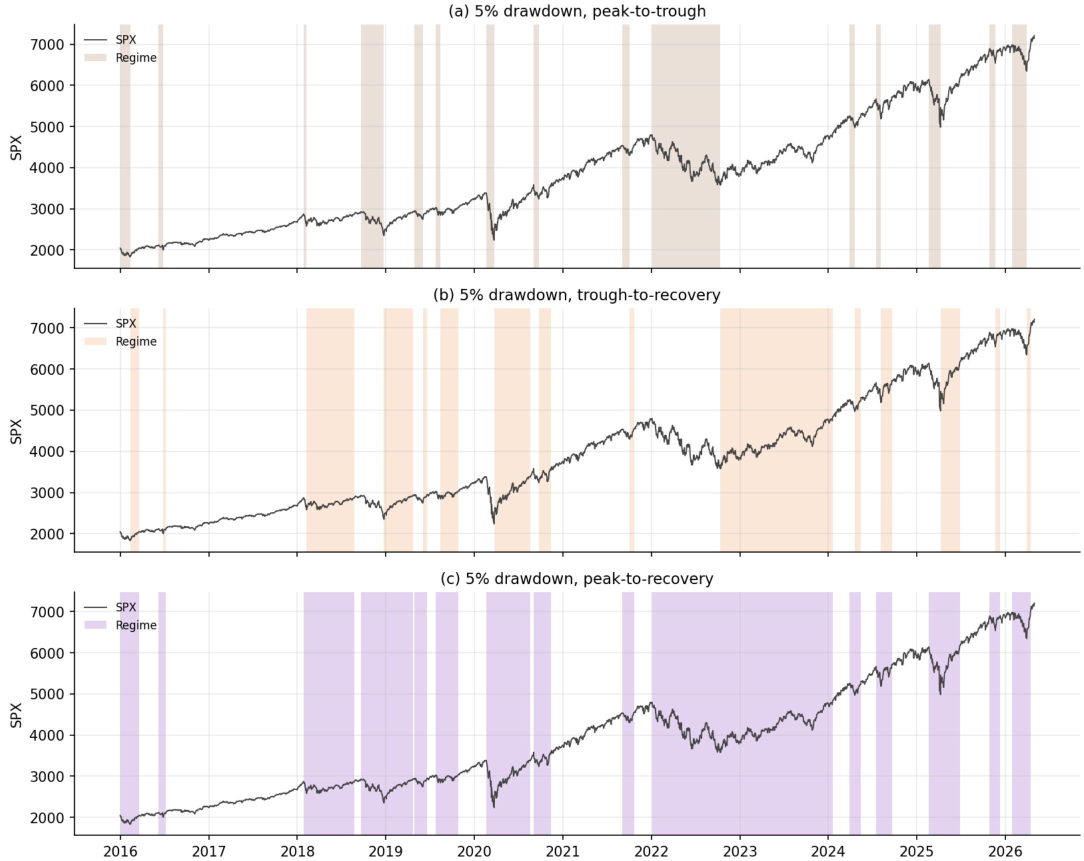
We re-estimate the regression from equation 6 in interaction form with HAC (Newey–West) standard errors at 10 lags and three pre-determined regime indicators D_t:

$$z_t = \alpha + \sum_{k=1}^{5} \beta_k \, z_{t-k} + \sum_{k=1}^{5} \big(\gamma_k + \delta_k \, D_t\big) \, R\text{-}RISI_{t-k} + \sum_{k=1}^{5} \psi_k \, VIX_{t-k} + \theta \, D_t + \phi' \, DOW_t + \varepsilon_t. \quad (8)$$

The results of the regime-amplification test are reported in Figure 13. The left panel classifies contraction periods as peak-to-trough, the middle panel shows the trough-to-recovery, and the right panel the peak-to-recovery specification. The contraction coefficient is the interaction between R-RISI and the regimes in (a)–(c). The expansion coefficient captures the R-RISI price impact outside these regimes. During the crash phase, the coefficient for the first lag of R-RISI is 3.4× larger inside an episode of market drawdown than outside (+0.0375 vs. +0.0109). The joint Wald test across all five lags does not reject the null, which shows no joint impact of all lags, consistent with the amplification being concentrated at lag 1. The asymmetry is therefore concentrated almost entirely at lag 1 and is diluted at the higher lags. The counterfactual in the middle panel, the trough-to-recovery, supports the finding of asymmetric price impacts during recessions. Here, the standalone (expansion coefficient) R-RISI term captures all non-recovery periods and is more pronounced and significant at lag 1. By contrast, the interaction between R-RISI and the recovery phase is not statistically significant and is of similar or smaller magnitude. The full peak-to-recovery period shows patterns similar to panel 1.

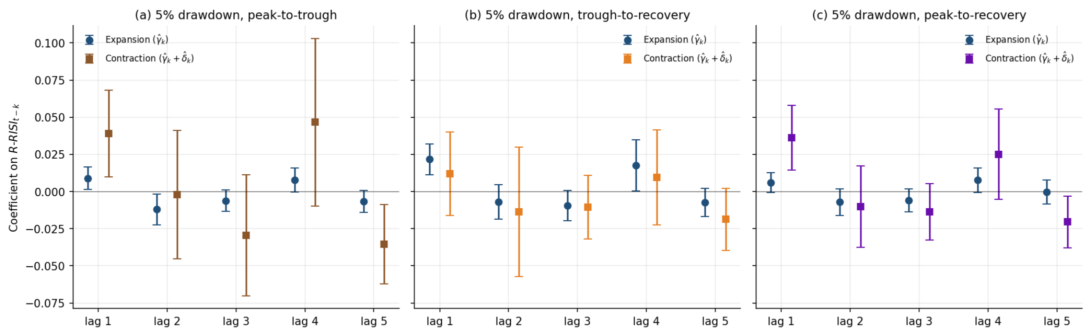
Overall, R-RISI carries stronger price-predictive content during market declines. Across regime indicators changes in R-RISI matter more for next-day SPX prices in downturns, consistent with García, 2013's recession-amplification result for news content.

Figure 12: S&P 500 with contraction-regime shading, three pre-determined indicators



Note: SPX time series with shaded contraction periods under the three regime indicators: peak-to-trough, trough-to-recovery and peak-to-recovery. The indicators decompose the drawdown cycle into its constituent phases using ex-post episode identification from 5 percent drawdown events in the SPX price series: peak-to-trough marks the crash phase, trough-to-recovery marks the recovery phase, and peak-to-recovery marks the full underwater period.

Figure 13: R-RISI lag coefficients on SPX_z, expansion vs. contraction (95% HAC CI)



Note: Lag coefficients $\hat{\gamma}_k$ (expansion baseline) and $\hat{\gamma}_k + \hat{\delta}_k$ (contraction) with 95% HAC confidence intervals under each of the three regime indicators in Equation (8).

7 Conclusion

R-RISI offers a new approach to measuring retail investor sentiment by using very granular Reddit data in combination with modern large language models. The indicator captures sentiment for a speculative retail investor cohort at very high-frequency and enables analyses on custom portfolios of stocks or crypto values. R-RISI is highly correlated with market prices and alternative retail sentiment measures. We show that R-RISI compares well towards existing indicators of retail investor activity, with the value-added of additional granularity, higher frequency and more flexible design. It reflects global events and structural market shifts in real time and is particularly informative during periods of market volatility. The descriptive properties of R-RISI offer a new lens to study retail investor behaviour and market sentiment to a novel degree. Its granularity shows how retail investors' sentiment and focus evolve over time and allows decomposition into topic- and asset-specific components down to individual securities.

We show the validity of R-RISI as a measure of retail investor activity empirically. Controlling for market factors, R-RISI has a signifcant impact on stock prices as well as crypto prices in the short-run. This effect is mostly driven by retail investor sentiment during contractionary market periods. This result allows twofold interpretation: (a) the statistically significant effect of R-RISI on market prices the next day establishes R-RISI's ability to capture real market activity and (b) the fact that this effect is stronger during market downturns points to R-RISI's cohort representativeness of speculative retail investors that are most active during periods of high market volatility. The latter is an important finding as retail investor behaviour, as captured by existing indices such as Stax and AAII, mostly reflects the passive investor flows of household savings, while R-RISI can capture the speculative flows of a mass of retail investors that has emerged with a wave of retail participation in the stock market over the last decade.

For policy makers and market observers, the R-RISI offers a scalable and dynamic tool to monitor shifts in retail sentiment across assets and sectors. Its timeliness and granularity make it well suited for use in financial stability monitoring, early-warning systems, and investor behaviour analysis by central banks, regulators, or other authorities.

Future work could study intraday sentiment around major news events, extend R-RISI to non-English, regional and asset-specific subreddits, decompose it into topical drivers using methods such as Principal Component Analysis, and examine links to retail trading activity.

References

AAII. (2026). *American association of individual investors: Investor sentiment survey*. Retrieved June 30, 2026, from https://www.aaii.com/sentimentsurvey.

Affuso, E., & Lahtinen, K. D. (2019). Social media sentiment and market behavior. *Empirical Economics*, *57*, 105–127. doi: 10.1007/s00181-018-1430-y.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294. doi: 10.1111/j.1540-6261.2004.00662.x.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. Retrieved June 30, 2026, from https://arxiv.org/abs/1908.10063.

Aubin, C., & Liedke, J. (2025, September 25). *Social media and news fact sheet*. Pew Research Center. Retrieved June 30, 2026, from https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680. doi: 10.1111/j.1540-6261.2006.00885.x.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–152. doi: 10.1257/jep.21.2.129.

Baklanova, V. (2025). The relationships between RedditSI and BTC exchange characteristics: Do reddit users still control the market? *Eurasian Economic Review*, *15*, 285–306. doi: 10.1007/s40822-024-00304-9.

Ballinari, D., & Behrendt, S. (2021). How to gauge investor behavior? a comparison of online investor sentiment measures. *Digital Finance*, *3*, 169–204. doi: 10.1007/s42521-021-00038-2.

Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention-induced trading and returns: Evidence from robinhood users. *The Journal of Finance*, *77*(6), 3141–3190. doi: 10.1111/jofi.13183.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 830–839. doi: 10.1609/icwsm.v14i1.7347.

Bloomberg Finance L.P. (2026). Bloomberg terminal data [Data retrieved using Bloomberg Terminal, April 2026].

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. doi: 10.1016/j.jocs.2010.12.007.

Charles Schwab. (2026). *Schwab trading activity index™ (STAX)*. Retrieved June 30, 2026, from https://www.schwab.com/investment-research/stax.

Chen, S., Peng, L., & Zhou, D. (2026). Wisdom or whims? decoding retail strategies with social media and AI. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4867401.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, *28*(3), 591–605. doi: 10.2307/1910133.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS: Investor sentiment and asset prices. *The Review of Financial Studies*, *28*(1), 1–32. doi: 10.1093/rfs/hhu072.

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, *66*(1), 67–97. doi: 10.1111/j.1540-6261.2010.01626.x.

European Central Bank. (2025, November). *Financial stability review - box: The role of household investors in market downturns* [November 2025 edition]. Retrieved June 30, 2026, from https://www.ecb.europa.eu/press/financial-stability-publications/fsr/focus/2025/html/ecb.fsrbox202511_02~3fb190ec50.en.html.

García, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267–1300. doi: 10.1111/jofi.12027.

Hernandez, J. P. (2025, April 28). *Reddit statistics in 2025 and tactics to grow your brand*. Sprout Social. Retrieved June 30, 2026, from https://sproutsocial.com/insights/reddit-statistics/.

Huang, C. C., & Shum, P. N. (2025). Dumb money? social network attention herding, sentiment, and markets. *The Journal of Finance and Data Science*, *11*, 100169. doi: 10.1016/j.jfds.2025.100169.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, *27*(4), 1116–1127. doi: 10.1016/j.ijforecast.2010.11.001.

Li, J., & Li, Z. (2026). Sentiment, social media and meme stock return predictability [Ahead of print]. *Review of Behavioral Finance*. doi: 10.1108/RBF-05-2025-0212.

Liu, Q., & Son, H. (2024). Methods for aggregating investor sentiment from social media. *Humanities and Social Sciences Communications*, *11*, 925. doi: 10.1057/s41599-024-03434-2.

Long, S. C., Lucey, B. M., Yarovaya, L., & Xie, Y. (2022). "I Just Like the Stock": The role of reddit sentiment in the gamestop share rally. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3822315.

Long, S. C., Xie, Y., Zhou, Z., Lucey, B., & Urquhart, A. (2025). From whales to waves: Social media sentiment, volatility, and whales in cryptocurrency markets. *The British Accounting Review*, 101682. doi: 10.1016/j.bar.2025.101682.

Mun, Y., & Kim, N. (2025). Leveraging large language models for sentiment analysis and investment strategy development in financial markets. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(2), 77. doi: 10.3390/jtaer20020077.

Münster, M., Reichenbach, F., & Walther, M. (2024). Robinhood, reddit, and the news: The impact of traditional and social media on retail investor trading. *Journal of Financial Markets*, *71*, 100929. doi: 10.1016/j.finmar.2024.100929.

Nguyen, B. T., Chu, T. T., Ha, S. X., & Nguyen, A. T. (2025). Decoding market sentiment: The power of ChatGPT in explaining Bitcoin returns from X data. *China Finance Review International*. doi: 10.1108/CFRI-05-2024-0278.

OpenAI. (2026). *GPT-5.1 and GPT-4o*. Retrieved June 30, 2026, from https://platform.openai.com/docs/models/.

RaiderBDev. (2026). *Reddit dataset*. Retrieved June 30, 2026, from https://academictorrents.com/browse.php?search=reddit.

Reichenbach, F., & Walther, M. (2023). Financial recommendations on reddit, stock returns and cumulative prospect theory. *Digital Finance*, *5*, 421–448. doi: 10.1007/s42521-023-00084-y.

Reuters. (2025). *Retail bought stocks at largest level over past decade, jpmorgan says*. Retrieved June 30, 2026, from https://www.reuters.com/markets/retail-bought-stocks-largest-level-over-past-decade-jpmorgan-says-2025-04-04/.

Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P.-M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862. doi: 10.1016/j.eswa.2023.119862.

Seetharam, Y., & Nyakurukwa, K. (2024). Headlines or hashtags? the battle in social media for investor sentiment in the stock market. *International Journal of Information Management Data Insights*, *4*(2), 100273. doi: 10.1016/j.jjimei.2024.100273.

Semenova, V., & Winkler, J. (2025). Social contagion and asset prices: Reddit's self-organized bull runs. *Quantitative Finance*, *25*(12), 1873–1904. doi: 10.1080/14697688.2025.2559970.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. doi: 10.1111/j.1540-6261.2007.01232.x.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437–1467. doi: 10.1111/j.1540-6261.2008.01362.x.

Wang, X., Xiang, Z., Xu, W., & Yuan, P. (2022). The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum. *Economics Letters*, *216*, 110598. doi: 10.1016/j.econlet.2022.110598.

A Appendix: Academic Torrents Dataset Files

The following torrent file names correspond to the exact datasets we used from the Academic Torrents web-based data storage platform. Searching online for these file names should lead directly to the dataset pages on Academic Torrents (RaiderBDev, 2026).

- 2005-06 to 2025-12: `reddit-3d426c47c767d40f82c7ef0f47c3acacedd2bf44.torrent`
- 2026-01: `reddit-8412b89151101d88c915334c45d9c223169a1a60.torrent`
- 2026-02: `reddit-c5ba00048236b60f819dbf010e9034d24fc291fb.torrent`
- 2026-03: `reddit-668087bb8c8c9c763b27a1a4c5e7fcb6add25f2c.torrent`
- 2026-04: `reddit-85d017ddd06920534187e7d45f21c7cec90c9bca.torrent`

B Appendix: Some Descriptives of the Reddit Post Raw Data

Table 8: Descriptives: summary table for selected variables

Subreddit	Variable	Period	Min	Q25	Mean	Median	Q75	Max	Stdev	Count
wallstreetbets	daily_posts	all	1	130	632.94	263	509	16993	4060.2	4126
wallstreetbets	daily_posts	excl_2021	1	119	326.37	232	428	5110	338.35	3761
wallstreetbets	score	all	0	1	105.9	1	2	41726	783.51	2611494
wallstreetbets	score	excl_2021	0	1	112.8	1	6	22946	1229	1227467
wallstreetbets	num_comments	all	0	0	34.14	1	3	10002	711.73	2611494
wallstreetbets	num_comments	excl_2021	0	0	49.87	2	9	10002	727.9	1227467
wallstreetbets	title_word_count	all	1	4	8.96	7	11	112	7.86	2611494
wallstreetbets	title_word_count	excl_2021	1	4	8.42	7	11	82	6.77	1227467
wallstreetbets	body_word_count	all	0	0	13.66	1	1	7079	96.97	2611494
wallstreetbets	body_word_count	excl_2021	0	0	22.05	1	1	7079	117.21	1227467
stocks	daily_posts	all	1	39	100.17	72	123	2493	125.8	4138
stocks	daily_posts	excl_2021	1	37	83.26	67	110	572	66.69	3773
stocks	score	all	0	1	33.69	1	2	101741	473.22	414512
stocks	score	excl_2021	0	1	32.74	1	3	5015	400.73	314145
stocks	num_comments	all	0	0	17.91	1	8	6798	86.54	414512
stocks	num_comments	excl_2021	0	0	18.28	1	9	4798	79.46	314145
stocks	title_word_count	all	1	4	8.26	7	11	67	5.71	414512
stocks	title_word_count	excl_2021	1	4	8.34	7	11	67	5.81	314145
stocks	body_word_count	all	0	1	54.62	1	53	6900	188.63	414512
stocks	body_word_count	excl_2021	0	1	57.66	1	58	6900	188.56	314145
StockMarket	daily_posts	all	1	15	45.07	28	60	1254	63.72	4136
StockMarket	daily_posts	excl_2021	1	14	34.9	25	50	393	29.11	3771
StockMarket	score	all	0	1	62.38	1	3	82143	729.61	186424
StockMarket	score	excl_2021	0	1	69.64	1	4	8214	826.4	131611
StockMarket	num_comments	all	0	0	13.43	0	5	8065	84.93	186424
StockMarket	num_comments	excl_2021	0	0	15.6	0	6	8065	97.4	131611
StockMarket	title_word_count	all	1	5	9.84	8	12	68	7.92	186424
StockMarket	title_word_count	excl_2021	1	5	9.91	8	12	65	7.49	131611
StockMarket	body_word_count	all	0	0	56.31	1	1	6793	252.36	186424
StockMarket	body_word_count	excl_2021	0	0	59.69	1	22	6793	253.95	131611
investing	daily_posts	all	13	61	92.73	79	104	1255	69.28	4138
investing	daily_posts	excl_2021	13	59	85.37	77	99	538	42.96	3773
investing	score	all	0	1	20.08	1	2	26836	167.62	383723

Continued on next page

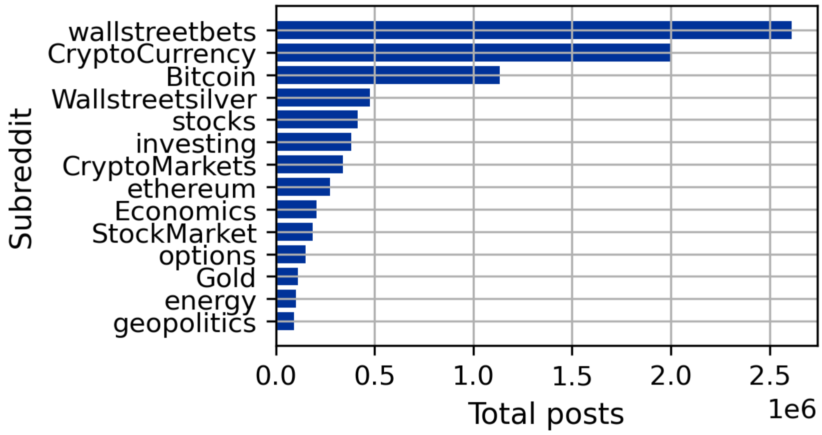
Subreddit	Variable	Period	Min	Q25	Mean	Median	Q75	Max	Stdev	Count
investing	score	excl_2021	0	1	20.06	1	2	14563	155.19	322093
investing	num_comments	all	0	1	15.03	2	9	5507	55.43	383723
investing	num_comments	excl_2021	0	1	15.59	2	11	2600	52.07	322093
investing	title_word_count	all	1	6	9.73	8	12	68	6.53	383723
investing	title_word_count	excl_2021	1	6	9.72	8	12	68	6.49	322093
investing	body_word_count	all	0	1	49.7	1	61	6279	131.82	383723
investing	body_word_count	excl_2021	0	1	54.3	1	70	6250	128.58	322093
ethereum	daily_posts	all	1	31	66.56	54	88	518	52.2	4136
ethereum	daily_posts	excl_2021	1	29	58.71	50	76	406	44.51	3771
ethereum	score	all	0	1	16.75	1	4	7415	111.66	275288
ethereum	score	excl_2021	0	1	14.21	1	5	7397	74.27	221381
ethereum	num_comments	all	0	0	6.51	0	3	4651	30.74	275288
ethereum	num_comments	excl_2021	0	0	6.04	0	3	4651	28.08	221381
ethereum	title_word_count	all	1	5	10.23	9	13	114	7.6	275288
ethereum	title_word_count	excl_2021	1	6	10.27	9	13	114	7.48	221381
ethereum	body_word_count	all	0	0	21.62	1	1	5769	93.71	275288
ethereum	body_word_count	excl_2021	0	0	23.19	1	1	5769	96.65	221381
CryptoMarkets	daily_posts	all	1	50	87.93	77	111	781	74.92	3876
CryptoMarkets	daily_posts	excl_2021	1	47	73.3	72	98	581	48.79	3511
CryptoMarkets	score	all	0	1	8.14	1	2	6984	60.86	340803
CryptoMarkets	score	excl_2021	0	1	6.73	1	2	4296	38.1	257339
CryptoMarkets	num_comments	all	0	0	3.67	0	1	3085	19.24	340803
CryptoMarkets	num_comments	excl_2021	0	0	3.81	0	1	1563	18.63	257339
CryptoMarkets	title_word_count	all	0	6	10.72	9	13	67	8.29	340803
CryptoMarkets	title_word_count	excl_2021	0	6	10.39	9	13	67	7.86	257339
CryptoMarkets	body_word_count	all	0	0	17.34	1	1	4656	78.65	340803
CryptoMarkets	body_word_count	excl_2021	0	0	19.48	1	1	4656	79.92	257339
CryptoCurrency	daily_posts	all	1	165	482.43	318	591	4423	542.54	4138
CryptoCurrency	daily_posts	excl_2021	1	148	360.94	279	511	3680	339.71	3773
CryptoCurrency	score	all	0	1	22.08	1	2	53679	292.78	1996309
CryptoCurrency	score	excl_2021	0	1	20.33	1	2	53679	243.8	1361824
CryptoCurrency	num_comments	all	0	0	18.58	1	6	59990	320.58	1996309
CryptoCurrency	num_comments	excl_2021	0	0	14.26	1	4	25360	132.73	1361824
CryptoCurrency	title_word_count	all	1	5	10.42	9	13	121	7.81	1996309
CryptoCurrency	title_word_count	excl_2021	1	6	10.46	9	13	89	7.65	1361824
CryptoCurrency	body_word_count	all	0	0	25.17	1	1	6807	107.2	1996309
CryptoCurrency	body_word_count	excl_2021	0	0	21.84	1	1	6686	100.92	1361824
Bitcoin	daily_posts	all	8	162	274.2	212	299	4191	231.11	4138
Bitcoin	daily_posts	excl_2021	8	159	258.31	204	279	4191	223.5	3773
Bitcoin	score	all	0	1	43.95	1	5	52169	303.91	1134629
Bitcoin	score	excl_2021	0	1	32.78	1	5	52169	283.42	974621
Bitcoin	num_comments	all	-2	0	12.57	2	8	8900	56.83	1134629
Bitcoin	num_comments	excl_2021	-2	0	12.17	2	8	7127	47.96	974621
Bitcoin	title_word_count	all	1	5	9.92	8	12	89	7.74	1134629
Bitcoin	title_word_count	excl_2021	1	5	9.89	8	12	89	7.57	974621
Bitcoin	body_word_count	all	0	0	26.26	1	14	6833	89.52	1134629
Bitcoin	body_word_count	excl_2021	0	0	26.62	1	16	6833	89.52	974621
Economics	daily_posts	all	3	30	49.93	41	57	523	34.99	4133
Economics	daily_posts	excl_2021	3	31	51.35	41	59	523	36.11	3768
Economics	score	all	0	1	133.15	1	13	5534	876.41	206345
Economics	score	excl_2021	0	1	131.2	1	12	5534	869.62	193469
Economics	num_comments	all	0	0	24.77	1	5	5180	105.24	206345
Economics	num_comments	excl_2021	0	0	24.2	1	4	5180	103.33	193469
Economics	title_word_count	all	1	8	11.91	10	13	66	7.29	206345
Economics	title_word_count	excl_2021	1	8	11.85	10	13	66	7.2	193469
Economics	body_word_count	all	0	0	0.86	0	0	5628	24.32	206345
Economics	body_word_count	excl_2021	0	0	0.9	0	0	5628	25.09	193469

Continued on next page

Subreddit	Variable	Period	Min	Q25	Mean	Median	Q75	Max	Stdev	Count
geopolitics	daily_posts	all	2	14	22.1	19	26	216	16.86	4138
geopolitics	daily_posts	excl_2021	2	14	22.58	19	26	216	17.43	3773
geopolitics	score	all	0	1	43.83	1	20	13921	143.7	91444
geopolitics	score	excl_2021	0	1	41.97	1	20	13921	137.3	85190
geopolitics	num_comments	all	0	0	17.65	1	12	4128	49.9	91444
geopolitics	num_comments	excl_2021	0	0	17.5	1	12	4128	49.05	85190
geopolitics	title_word_count	all	1	8	11.63	10	14	62	6.9	91444
geopolitics	title_word_count	excl_2021	1	8	11.62	10	14	62	6.84	85190
geopolitics	body_word_count	all	0	0	21.54	0	1	6339	135.03	91444
geopolitics	body_word_count	excl_2021	0	0	21.38	0	1	5744	132.34	85190
energy	daily_posts	all	2	17	24.96	25	31	128	10.57	4138
energy	daily_posts	excl_2021	2	17	25.09	25	31	128	10.77	3773
energy	score	all	0	1	39	2	16	47403	294.55	103286
energy	score	excl_2021	0	1	39.58	2	15	47403	306.94	94646
energy	num_comments	all	0	0	12.27	1	7	5843	67.34	103286
energy	num_comments	excl_2021	0	0	12.22	1	7	5843	69.68	94646
energy	title_word_count	all	1	8	14.34	11	15	66	10.65	103286
energy	title_word_count	excl_2021	1	8	14.13	11	15	66	10.45	94646
energy	body_word_count	all	0	0	12.96	0	0	4593	76.61	103286
energy	body_word_count	excl_2021	0	0	13.21	0	0	4593	77.93	94646
Gold	daily_posts	all	1	5	27.65	21	40	363	28.48	4045
Gold	daily_posts	excl_2021	1	5	27.97	20	42	363	29.76	3680
Gold	score	all	0	1	30.42	2	15	55236	218.18	111832
Gold	score	excl_2021	0	1	30.55	3	14	55236	226.81	102920
Gold	num_comments	all	0	0	11.93	3	12	5953	33.31	111832
Gold	num_comments	excl_2021	0	0	12.03	3	12	5953	34.4	102920
Gold	title_word_count	all	1	4	8.86	6	11	68	7.78	111832
Gold	title_word_count	excl_2021	1	4	8.66	6	11	66	7.56	102920
Gold	body_word_count	all	0	0	22.18	1	25	6804	63.35	111832
Gold	body_word_count	excl_2021	0	0	22.67	1	26	6804	61.63	102920
Wallstreetsilver	daily_posts	all	9	32	248.59	78	411	2317	301.77	1918
Wallstreetsilver	daily_posts	excl_2021	9	29	156.11	49	278	963	180.44	1581
Wallstreetsilver	score	all	0	23	86.03	48	95	2660	138.38	476802
Wallstreetsilver	score	excl_2021	0	22	75.85	47	89	1513	108.9	246807
Wallstreetsilver	num_comments	all	0	1	11.16	4	10	10001	24.3	476802
Wallstreetsilver	num_comments	excl_2021	0	1	11.01	4	10	15413	55.63	246807
Wallstreetsilver	title_word_count	all	1	5	12.86	9	16	95	11.3	476802
Wallstreetsilver	title_word_count	excl_2021	1	5	13.23	10	17	67	11.55	246807
Wallstreetsilver	body_word_count	all	0	0	22.34	0	1	7106	107.68	476802
Wallstreetsilver	body_word_count	excl_2021	0	0	18.53	0	1	6637	91.07	246807
options	daily_posts	all	1	11	37.04	26	49	718	42.29	4113
options	daily_posts	excl_2021	1	10	30.95	24	43	718	30.99	3748
options	score	all	0	1	10.43	1	3	31625	110.47	152335
options	score	excl_2021	0	1	8.77	1	3	6452	52.31	115998
options	num_comments	all	0	1	11.46	2	10	3240	37.75	152335
options	num_comments	excl_2021	0	1	10.9	2	10	3240	30.69	115998
options	title_word_count	all	1	4	8.25	7	10	68	6.53	152335
options	title_word_count	excl_2021	1	4	8.03	7	10	68	6.02	115998
options	body_word_count	all	0	1	57.33	1	69	5877	138	152335
options	body_word_count	excl_2021	0	1	58.01	1	71	5655	133.33	115998

Note: The table reports descriptive statistics for daily Reddit post counts, scores, comment numbers, and title word counts by subreddit. Values are shown for the full sample and for the sample excluding 2021. Mean, percentiles (Q25, Median, Q75), minimum, maximum, and standard deviation are reported. Count is the number of days or posts.

Figure 14: Subreddits Ranked by Aggregated Post Count

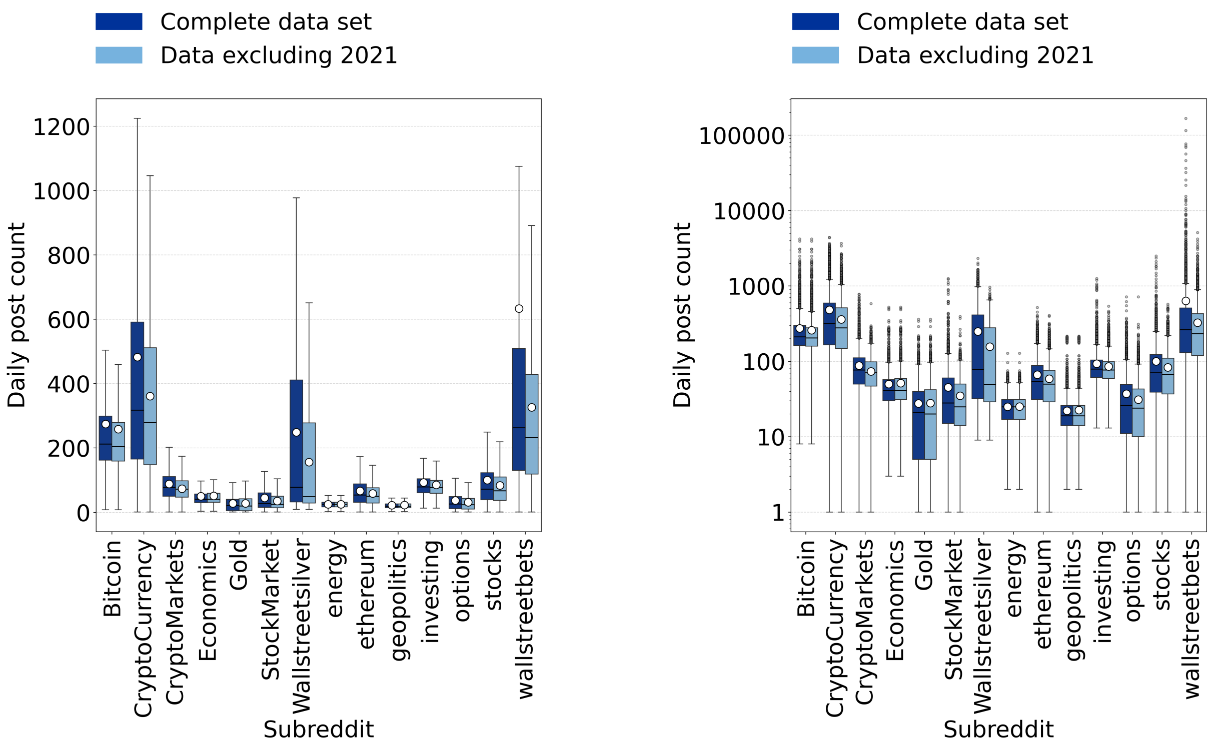


Note: Subreddits are ranked by their aggregated post count (full data period). X-axis values are in millions.

Figure 15: Distribution of daily Reddit posts per subreddit

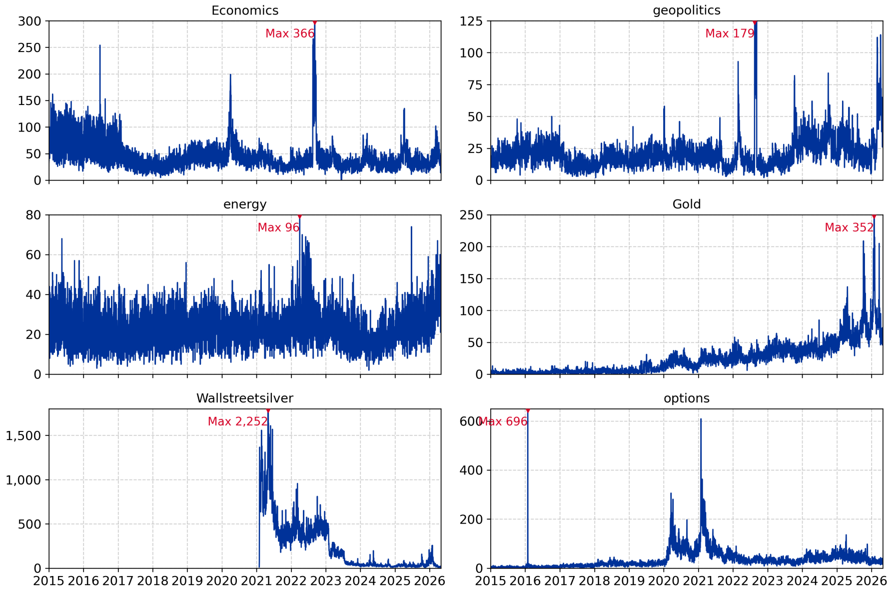
Panel A: Boxplot not showing outliers

Panel B: Boxplot showing outliers



Note: Panel A shows boxplots of daily Reddit post counts per subreddit without outliers, while Panel B includes them. Boxes represent the interquartile range (25th–75th percentile), with the horizontal line marking the median and the grey dot the mean. Whiskers extend to 1.5 × IQR from the box, and points beyond indicate outliers. Dark blue boxes refer to the full sample; light blue boxes exclude 2021 observations.

Figure 16: Daily Posting Activity Across the Remaining Subreddits



Note: The figure shows the daily number of Reddit posts between January 2015 and April 2026 for the remaining subreddits in the data set.

C Appendix: LLM AI Model Prompts

ChatGPT-4o

CLASSIFIER_SYSTEM_PROMPT = ”””You are an expert in classifying sentiment and topics from short texts. You will be given Reddit post titles from finance, investing, and crypto subreddits such as r/wallstreetbets.Financial jargon, Reddit slang, tickers (e.g. $AAPL, BTC), and option/commodity terms are common – interpret them precisely. Return ONLY valid JSON. No markdown fences. Do not include any text before or after.”””

USER_PROMT = f””” You are a Reddit title post classification expert. You will extract sentiment and topic of each Reddit post.Be aware of Reddit user slang and the abbreviations they use and their meanings, e.g. FOMO, FUD, diamond hands(hold)/paper hands(sell to soon)/HODL(hold the line, do not sell), copium/hopium, too the Moon, rocket, 10x, Buy the dip (BTD), rug pull, pump-and-dump, wagmi/ngmi, gas war, YOLO (big, reckless trades), gain porn/loss porn, tendies, rekt (totally blown position), ape, whale, autist/degen, stonk, btfd (buy the fucking dip), bagholder, theta gang/gamma gang, IV crush, etc.

I will give you a list of Reddit post titles. Each row contains:

- ”row_id”: a unique identifier
- ”title_adjusted”: the post title (text)

Your task is to label each row with:

1. **”sentiment_label”** — Choose strictly ONE of the three options from: [”positive”, ”negative”, ”neutral”]:
- ”positive”: clear optimism, enthusiasm, or bullish tone, positve emojies
- ”negative”: clear concern, criticism, fear, or bearish tone, negative emojies
- ”neutral”: factual, inquisitive, balanced, or uncertain

If both positive and negative cues appear, pick the dominant tone; if dominance is unclear, label neutral.

2. **”sentiment_confidence”** — A number between 0.0 and 1.0 reflecting how confident you are in the classification.

3. **”topic_label”** — Choose strictly ONE of the options from: [”crypto”, ”options”, ”funds”, ”commodities”, ”macro”, ”tariffs”, ”geopolitical”, ”equities”, ”other”]:
Use the following as guideline:
- ”crypto”: crypto currencies (e.g. bitcoin, BTC, $BTC) or tokens (BTC, ETH, SOL, DOGE, NFTs), stablecoin, on-chain, crypto markets, digital currencies).
- ”options”: calls, puts, IV, volatility, delta/gamma/vega, Greeks, spreads, LEAPS.
- ”funds”: ETFs, mutual funds, hedge funds, index trackers, closed-end funds (SPY, QQQ, ARKK, VTSAX...).
- ”commodities”: Physical or futures on oil, metals, gold, silver, wheat, copper, nat-gas, lithium...
- ”macro”: Economy-wide data or policy, rates, inflation, ECB/Fed decisions, CPI, GDP, unemployment, recession talk
- ”tariffs”: Import/export duties, trade restrictions, anti-dumping, section-301, retaliatory levies, protectionism related to international trade, trade wars
- ”geopolitical”: Political or military events _not centred on trade_, elections, wars, coups, sanctions **because** of conflict.
- ”equities”: All other stock-market investing, trading, finance, corporate news: tickers (e.g. $AAPL), profits, gains/losses made, earnings, upgrades/downgrades, e.g. ”panic sell”, “buy the dip”. (Default here **only if** none of the higher-priority labels above clearly apply.)
- ”other”: Use **only** when the post is **not** about finance/investing (memes, personal life, off-topic).
Rule of thumb:
- If the post is finance-related but ambiguous, walk down the priority list until the first definite match.
- A crypto ticker or crypto related overrides equities; options e.g. on TSLA overrides equities; a tariff headline on steel overrides commodities, etc.
OUTPUT: Return a valid JSON array ONLY (no extra text). Each element must be:
{ ”row_id”: <`input row_id`>, ”sentiment_label”: <`positive|negative|neutral`>, ”sentiment_confidence”: <`float 0.0-1.0`>, ”topic_label”: <`crypto | options | funds | commodities | macro | tariffs | geopolitical | equities | other`>}
... Rows: {rows_json} ”””.strip()

ChatGPT-5.1 (more compact due to reduce token usage and accuracy by avoiding overkill promt information)
CLASSIFIER_SYSTEM_PROMPT = ”””You are an expert in classifying sentiment, topics, and extracting referenced assets from short texts. You will be given Reddit post titles and text body from finance, investing, and crypto or commodities subreddits such as r/wallstreetbets and others. Financial jargon, community slang, tickers, crypto symbols, index symbols, and commodity references are common. Interpret them precisely. Return ONLY valid JSON. No markdown fences. Do not include any text before or after.”””
SAFE_SLANG_NOTE = ””” Be aware of common finance community slang and abbreviations, for example: FOMO, FUD, diamond hands, paper hands, HODL, copium, hopium, to the moon, rocket, 10x, buy the dip (BTD or BTFD), rug pull, pump and dump, wagmi or ngmi, gas war, YOLO, gain porn or loss porn, tendies, rekt, ape, whale, bagholder, theta gang, gamma gang, IV crush. ”””
USER_PROMT = f””” Classify each row. Return ONLY a JSON array. SAFE_SLANG_NOTE For each row return:
- sentiment_label: positive/negative/neutral
- sentiment_confidence: 0.0-1.0
- sentiment_strength: -100 to 100
- topic_label: crypto/options/funds/commodities/macro/tariffs/geopolitical/equities/other
- topic_confidence: 0.0-1.0
- free_topic: 2-5 word market narrative (no slang)
- tickers: list of uppercase standard trading symbols ([] if none)
- trade_direction: long/short/hedged/neutral/unclear
- market_emotion: panicked/fearful/cautious/neutral/optimistic/euphoric

- region: US/EU/Asia/Global/Other

- time_horizon: intraday/short_term/medium_term/long_term/unclear

Tickers: recognize names/slang (written out and ticker) → standard symbol, e.g. Google or Googl, Volkswagen or VOW3. Crypto: e.g. BTC,ETH,SOL, Bitcoin, Ethereum, stablecoins e.g USDT or Thether, etc. Indices: e.g. SPX,VIX, Sp500. commodities etc. Topics: if multiple mentioned, choose dominant. If unclear, use "other".

Return exactly len(rows) objects with matching row_id values. JSON array only, no markdown."""

D Appendix: List of AI tickers in the S&P500 index

Direct AI: NVIDIA, Microsoft, Apple, Alphabet, Amazon, Meta, Broadcom, Tesla, Oracle, Palantir, AMD, Salesforce, IBM, Uber, ServiceNow, Qualcomm, Arista, Adobe, Micron, Palo Alto, Intel, Crowdstrike, Cadence Design, Dell, NXP, Fortinet, Digital Reality Trust, HP and Super Micro Computer; *AI utilities:* NRG, Vistra, NextEra, Southern, Constellation, Public Service Enterprise, Entergy and NiSource; *AI capital equipment:* Eaton, Trane, Johnson Controls and Quanta

E Appendix: Consistency check outputs

This appendix reports additional results from the repeated-labeling exercise described in Section 4.1. The same set of 73,669 Reddit posts from December 2018 and December 2025 was independently classified ten times using the same GPT-5.1 prompt and model settings. Each repeated classification was then compared with the baseline classification.

Table 9: Internal Consistency of GPT-5.1 Classifications Across Ten Repeated Runs

Variable	**Rows**	**Mean matches out of 10**	**Agreement vs. baseline**	**All 10 match**	**Share all 10 match**	**Share zero match**
Sentiment label	73,669	8.89	88.9%	58,407	79.3%	4.6%
Topic label	73,669	9.11	91.1%	62,419	84.7%	4.3%
Tickers	73,669	9.14	91.4%	60,784	82.5%	3.1%
Region	73,669	8.77	87.7%	56,977	77.3%	4.6%
Trade direction	73,669	7.38	73.8%	38,047	51.6%	9.7%
Market emotion	73,669	7.96	79.6%	46,654	63.3%	8.4%
Time horizon	73,669	7.53	75.3%	42,385	57.5%	10.6%

Notes: The table reports internal consistency between the baseline GPT-5.1 classification and ten repeated classifications of the same posts. "Mean matches out of 10" denotes the average number of repeated runs that matched the baseline label for a given observation. "Agreement vs. baseline" divides this number by ten. "All 10 match" reports the number of observations for which all ten repeated classifications matched the baseline. Confidence scores are excluded from the table as they are continuous outputs.

Table 10: Distribution of Row-Level Consistency Counts Across Ten Repeated Runs

Variable	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sentiment label	4.6%	1.6%	1.3%	1.2%	1.2%	1.2%	1.5%	1.7%	2.3%	4.2%	79.3%
Topic label	4.3%	1.2%	0.9%	0.8%	0.8%	0.8%	1.0%	1.1%	1.6%	2.8%	84.7%
Tickers	3.1%	1.3%	1.0%	1.0%	1.0%	1.1%	1.2%	1.6%	2.1%	4.2%	82.5%
Region	4.6%	1.9%	1.5%	1.4%	1.4%	1.5%	1.7%	1.9%	2.6%	4.1%	77.3%
Trade direction	9.7%	4.1%	3.2%	3.1%	3.1%	3.3%	3.7%	4.1%	5.4%	8.7%	51.6%
Market emotion	8.4%	2.9%	2.4%	2.2%	2.2%	2.3%	2.5%	3.0%	4.0%	6.7%	63.3%
Time horizon	10.6%	3.8%	3.0%	2.6%	2.6%	2.7%	2.8%	3.2%	4.3%	7.0%	57.5%

Notes: Each row shows how often the repeated classifications matched the baseline classification for a given variable. The columns 0 to 10 indicate the number of repeated runs, out of ten, that reproduced the baseline label for the same post. For example, for the sentiment label, the value of 79.3% in column 10 means that 79.3% of posts received the same sentiment label as the baseline in all ten repeated runs. The value of 4.6% in column 0 means that, for 4.6% of posts, none of the ten repeated sentiment labels matched the baseline.

Table 11: Internal Consistency by Confidence-Based Sample Split

Sample	Variable	Rows	Mean matches out of 10	Agreement vs. baseline	All 10 match	Share all 10 match	Share zero match
High confidence	Sentiment label	51,324	8.98	89.8%	41,411	80.7%	4.1%
High confidence	Topic label	51,324	9.56	95.6%	47,060	91.7%	1.9%
High confidence	Tickers	51,324	8.93	89.3%	40,235	78.4%	3.9%
High confidence	Region	51,324	9.23	92.3%	43,803	85.3%	2.9%
High confidence	Trade direction	51,324	7.45	74.5%	27,359	53.3%	9.5%
High confidence	Market emotion	51,324	8.02	80.2%	32,768	63.8%	8.1%
High confidence	Time horizon	51,324	7.34	73.4%	27,873	54.3%	11.3%
Lower confidence	Sentiment label	22,345	8.66	86.6%	16,996	76.1%	5.6%
Lower confidence	Topic label	22,345	8.10	81.0%	15,359	68.7%	9.7%
Lower confidence	Tickers	22,345	9.62	96.2%	20,549	92.0%	1.2%
Lower confidence	Region	22,345	7.71	77.1%	13,174	59.0%	8.5%
Lower confidence	Trade direction	22,345	7.21	72.1%	10,688	47.8%	10.0%
Lower confidence	Market emotion	22,345	7.84	78.4%	13,886	62.1%	9.0%
Lower confidence	Time horizon	22,345	7.96	79.6%	14,512	64.9%	9.0%

Notes: The high-confidence sample includes posts with sentiment confidence of at least 0.7 and topic confidence of at least 0.7. The lower-confidence sample includes all remaining observations. The table reports agreement between the baseline GPT-5.1 classification and ten repeated classifications. Confidence scores are excluded from the table.

Table 12: Sentiment Label Transitions Between Baseline and Repeated Classifications

Baseline label	Repeated label	Count	Share of comparisons
Neutral	Neutral	293,957	39.9%
Positive	Positive	235,651	32.0%
Negative	Negative	124,945	17.0%
Positive	Neutral	29,668	4.0%
Neutral	Positive	21,956	3.0%
Negative	Neutral	15,693	2.1%
Neutral	Negative	11,127	1.5%
Positive	Negative	2,161	0.3%
Negative	Positive	1,522	0.2%

Notes: Shares are calculated over all baseline-versus-repeat comparisons, where each observation is compared separately with each of the ten repeated labeling runs. The denominator is therefore 736,690 comparisons. Direct polarity reversals refer to positive-to-negative and negative-to-positive switches. Neutral shifts refer to transitions between positive/negative and neutral in either direction.

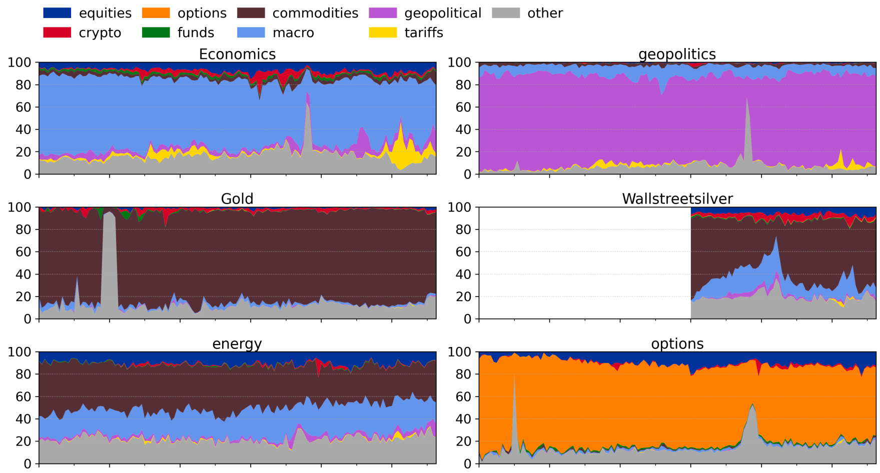
Table 13: Row-Level Sentiment Instability Across Ten Repeated Runs

Type of instability	Rows	Total rows	Share of rows
At least one positive-negative contradiction	807	73,669	1.1%
At least one neutral-related shift	14,787	73,669	20.1%

Notes: Row-level shares report the proportion of observations for which at least one of the ten repeated classifications differs from the baseline in the specified way. Positive-negative contradictions are cases where at least one repeated classification switches between positive and negative relative to the baseline. Neutral shifts are cases where at least one repeated classification switches between a directional label and neutral.

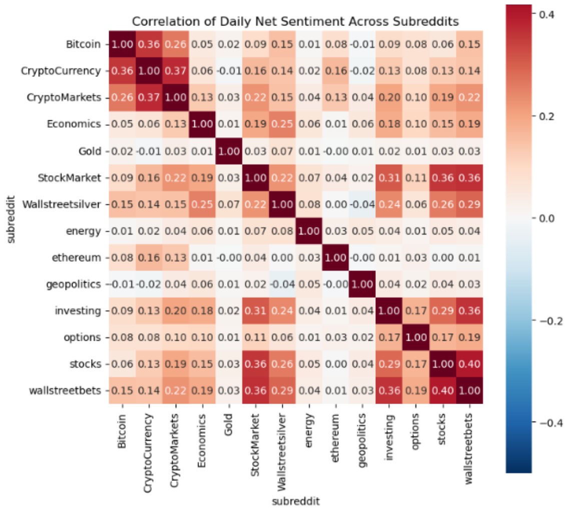
F Appendix: Additional Results

Figure 17: Reddit Discussion Topics Across Remaining Subreddits Over Time



Note: The chart depicts the share of Reddit posts per topic category aggregated at a monthly frequency. Topic categories were assigned as described in Section 4.1. Data are from January 2015 to April 2026.

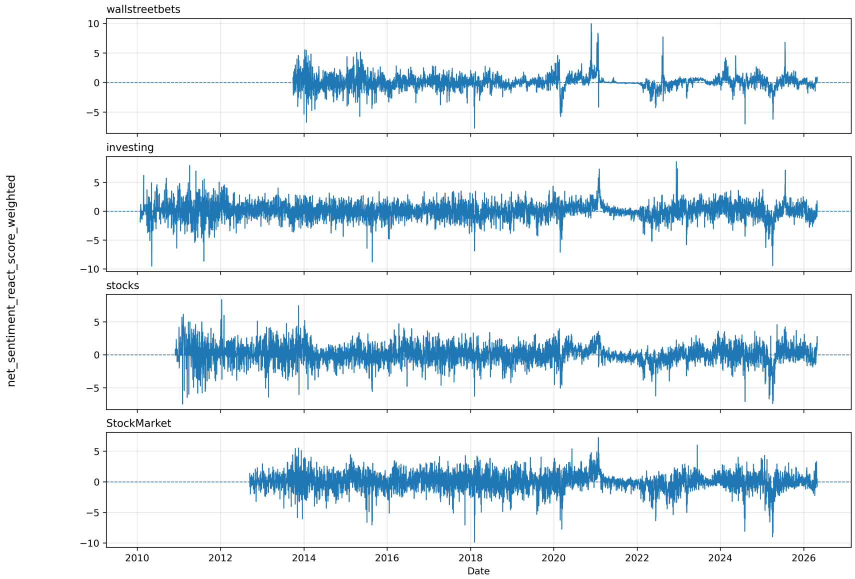
Figure 18: R-RISI Subreddit Correlation



Note: The chart shows the correlation of reactiveness-weighted net sentiment across subreddits.

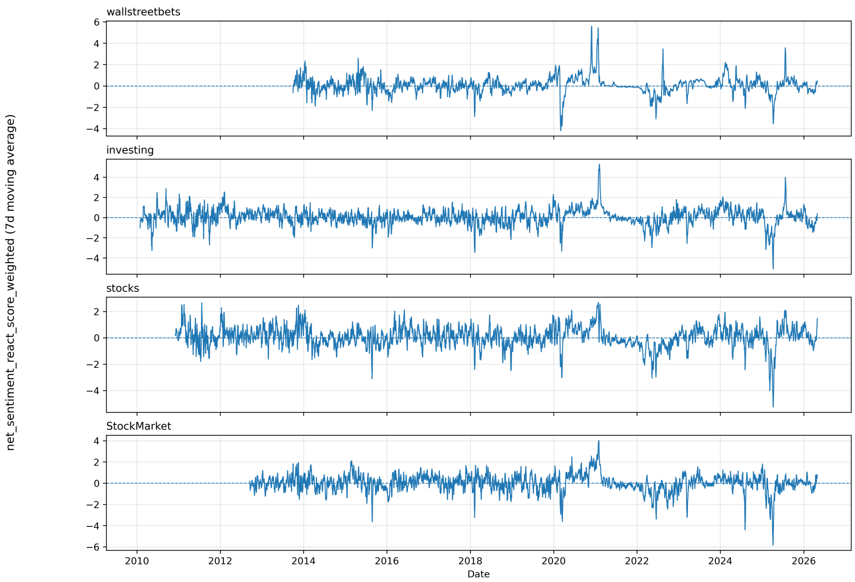
G Appendix: Sentiment per subreddit

Figure 19: Net sentiment investing subreddits - no smoothing



Note: net retail sentiment is constructed from Reddit post titles on from 2015 onwards. Posts are classified into market-related topics (equities, funds, macro, options, commodities, geopolitics, tariffs and crypto) using OpenAI's GPT-5.1 model, with unrelated content, such as community posts, excluded. The posts are labeled as positive, negative or neutral. Net sentiment is defined as the difference between the 365-day z-scores of positive and negative posts, where each post is weighted by the logarithm of its comments and upvotes to proxy its relative importance in the forum. This figure shows daily data unsmoothed for each of the extracted investing subreddits.

Figure 20: Net sentiment investing subreddits - 7 day smoothing

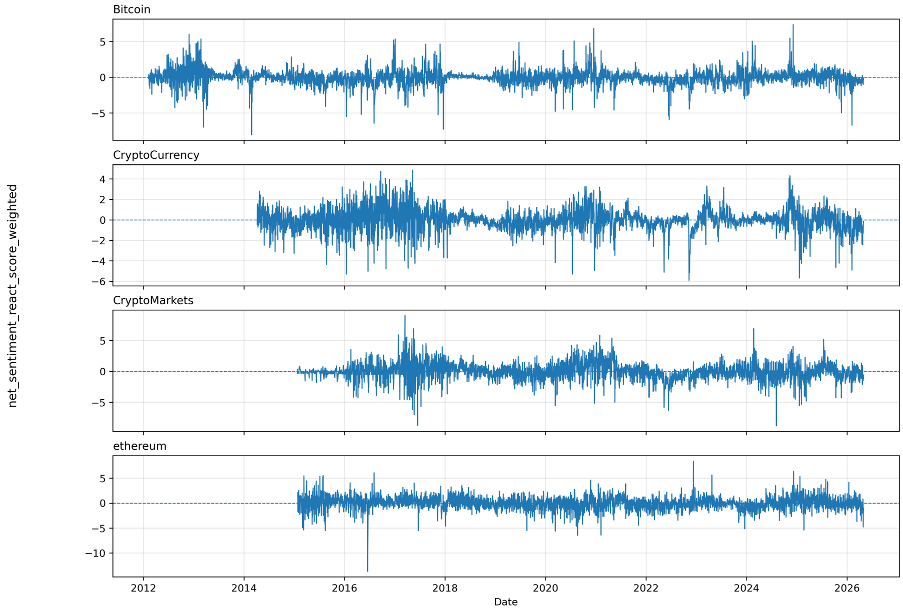


Note: See the prior note. Here, timeseries are smoothed by a rolling seven day mean.

Figure 21: Net sentiment investing subreddits - 30 day smoothing

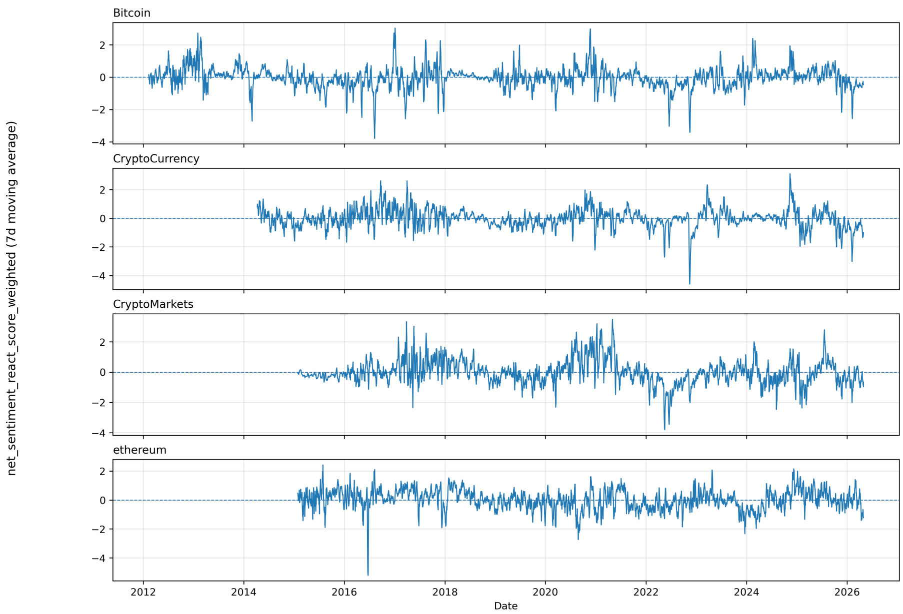
Note: See the prior note. Here, timeseries are smoothed by a rolling 30 day mean.

Figure 22: Net sentiment crypto subreddits - no smoothing



Note: See prior note. This figure shows the net sentiment unsmoothed for the extracted crypto subreddits.

Figure 23: Net sentiment crypto subreddits - 7 day smoothing

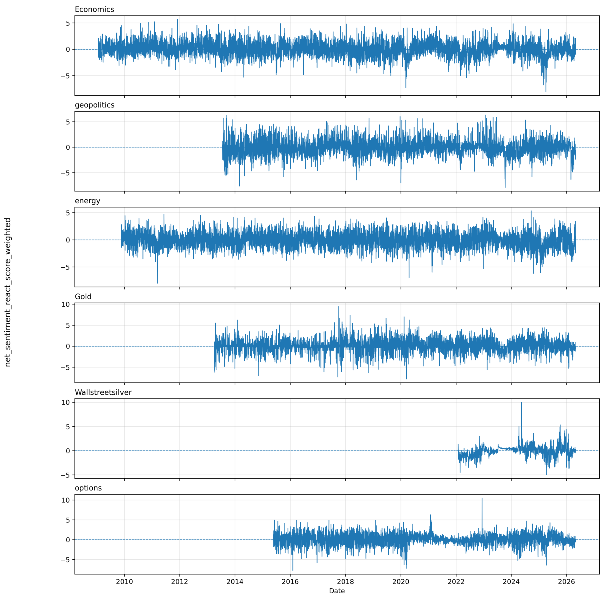


Note: See prior note. Here, the timeseries are smoothed by a seven day moving average.

Figure 24: Net sentiment crypto subreddits - 30 day smoothing

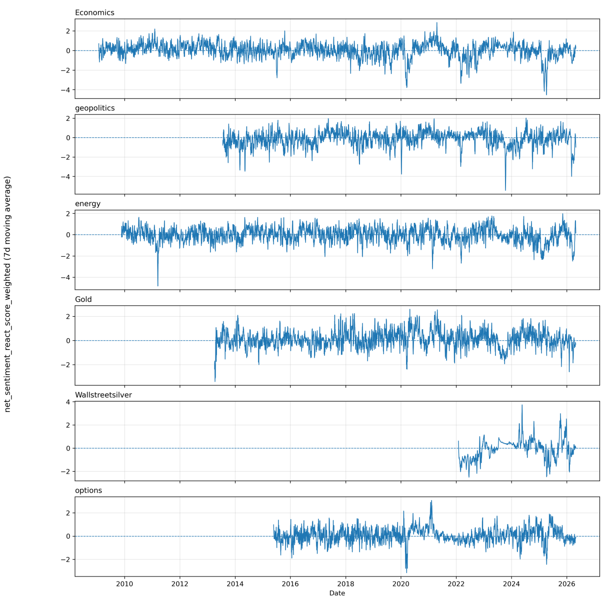
Note: See prior note. Here, the timeseries are smoothed by a 30 day moving average.

Figure 25: Net sentiment remaining subreddits - no smoothing



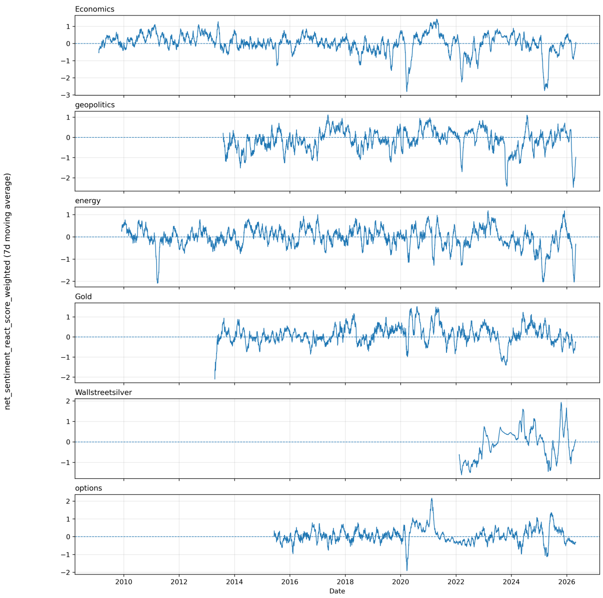
Note: See prior note. This figure shows the net sentiment unsmoothed for the extracted remaining subreddits.

Figure 26: Net sentiment remaining subreddits - 7 day smoothing



Note: See prior note. Here, the timeseries are smoothed by a seven day moving average.

Figure 27: Net sentiment remaining subreddits - 30 day smoothing



Note: See prior note. Here, the timeseries are smoothed by a 30 day moving average.

H Appendix: R-RISI and trading activity

This appendix provides additional evidence on the relationship between R-RISI and market prices of stock indices and Bitcoin, as well as total transaction volumes for the SPX index.

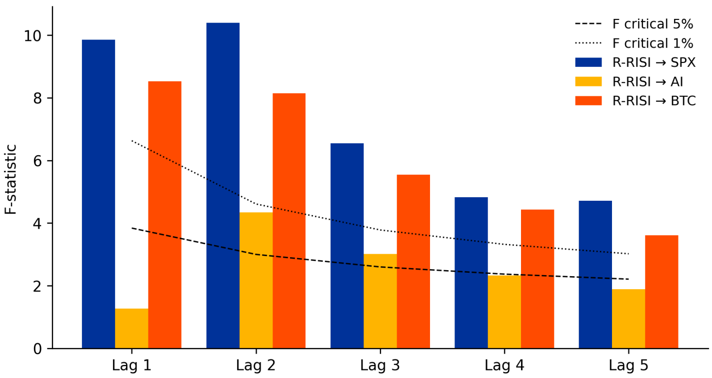
H.1 Additional results for the relationship between sentiment and market prices

Granger causalities of the first five lags in R-RISI on SPX and AI stock prices as well as the price on Bitcoin are reported in Figure 28. R-RISI granger–causes SPX stock prices at the 5% and 1% significance level. For a more granular sample on AI stocks, we cannot find granger-causality at the 1% significance–level, but at the 5%–level for lags 2–4. The influence of R-RISI expressing sentiment about Bitcoin on the price of Bitcoin itself is significant across all lags. The Granger-causality test therefore establishes a directional predicatbility from R-RISI to market prices in a bivariate regression.

H.2 unweighted R-RISI and market prices

There is valid concern that weighting Reddit posts by user engagement introduces a hindsight bias. In principle, users can react to a post even days after, which can cause the user engagement to not be related to the initial post alone, but also to market movements. Therefore, R-RISI is running the risk of being validated ex-post. To address this concern, we are running the regressions from equation 6 again but instead of using the weighted R-RISI, we are regressing market prices on the unweighted R-RISI, which does not contain the hinsight bias. Results are reported in Table 14. We show that our results are robust to using the unweighted version of R-RISI.

Figure 28: R-RISI and market prices: Granger causality



Note: F-statistics for Granger causality from R-RISI to market prices at lags 1–5. Horizontal dashed lines mark 5% and 1% critical values under the large-df $F(q, \infty)$ approximation.

Table 14: Effect of unweighted R-RISI on SPX, AIX and BTC

	SPX		AIX		BTC	
	γ_k	p-value	γ_k	p-value	γ_k	p-value
$Constant$	0.0243	0.210	0.0054	0.852	0.0000	0.997
z_{t-1}	0.8717	0.000	0.9497	0.000	0.9329	0.000
z_{t-2}	0.1426	0.125	0.0242	0.686	0.0893	0.125
z_{t-3}	-0.0573	0.385	0.0076	0.895	-0.0208	0.643
z_{t-4}	-0.0355	0.445	0.0205	0.696	-0.0001	0.998
z_{t-5}	0.0604	0.099	0.0212	0.705	-0.0073	0.833
$RISI_{t-1}$	0.0208	0.001	0.0101	0.195	0.0146	0.004
$RISI_{t-2}$	-0.0120	0.104	0.0012	0.839	-0.0035	0.539
$RISI_{t-3}$	-0.0090	0.088	0.0008	0.898	-0.0048	0.422
$RISI_{t-4}$	0.0180	0.038	0.0023	0.684	0.0003	0.955
$RISI_{t-5}$	-0.0093	0.046	0.0044	0.414	-0.0051	0.308
VIX_{t-1}	-0.0011	0.453	0.0011	0.387	0.0005	0.643
VIX_{t-2}	-0.0022	0.155	0.0020	0.410	-0.0008	0.388
VIX_{t-3}	0.0003	0.814	0.0044	0.154	0.0000	0.953
VIX_{t-4}	0.0000	0.935	0.0006	0.727	0.0016	0.048
VIX_{t-5}	0.0026	0.014	0.0000	0.958	-0.0010	0.155

Note: The table reports estimated coefficients for the model in Equation 6, replacing weighted R-RISI with unweighted R-RISI to address hindsight bias.

H.3 R-RISI and retail trading volume

A complementary test of the behavioural content of R-RISI examines its relationship to aggregate trading volume, which reflects realised market activity rather than stated sentiment. This subsection reports the contemporaneous and dynamic association between daily R-RISI and the logarithm of S&P 500 daily share volume (hereafter log SPX_VOL) over the period January 2016 to December 2025.[3]

The Pearson correlation between daily R-RISI and log SPX_VOL is -0.171, and the correlation between |R-RISI| and

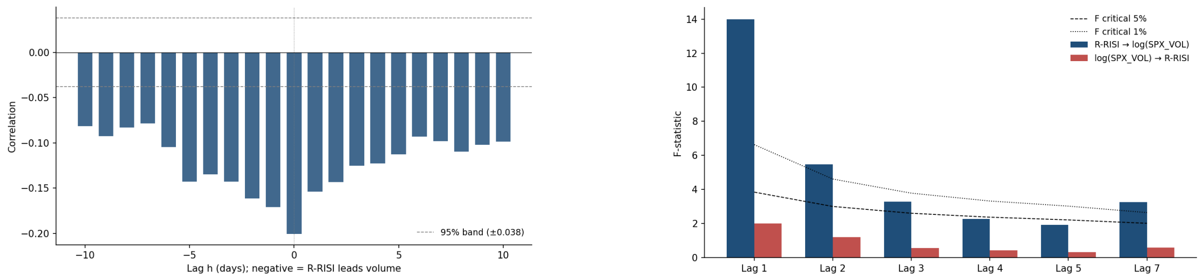
[3] SPX_VOL is an aggregate market measure rather than a retail-specific one. Separating retail execution from institutional flow would sharpen the interpretation but was not feasible because the necessary data is scarce and not available for this analysis.

log SPX_VOL is +0.121 ($p < 0.0001$). The negative first-moment correlation indicates that bullish sentiment tends to coincide with lower trading activity, while the positive second-moment correlation indicates that extreme sentiment in either direction is associated with elevated volume. The cross-correlation function (Figure 29, Panel A) is uniformly negative across lags −10 to +10 trading days, peaking at $h = 0$ ($\rho = -0.171$), with mild asymmetry suggesting that R-RISI leads log volume slightly rather than the reverse. Bivariate Granger-causality tests from R-RISI to log SPX_VOL are significant at every lag from 1 to 5 (Figure 29, Panel B).

Figure 29: R-RISI and Log S&P 500 Share Volume: Cross-Correlation Function and Granger causality

Panel A: CCF, ±10 days

Panel B: Granger Causality: F-Statistics



Note: Panel A shows the cross-correlation function $\rho(h) = \text{corr}(\text{R-RISI}_{t+h}, \log \text{SPX\_VOL}_t)$ for $h \in \{-10, \ldots, +10\}$ trading days. Dashed lines are approximate 5% significance bands. Panel B shows F-statistics for Granger causality from R-RISI to log SPX_VOL (blue bars) and from log SPX_VOL to R-RISI (red bars) at lags 1–5. Horizontal dashed lines mark 5% and 1% critical values.

To assess whether the sentiment–volume association holds when testing for predictive power, while controlling for volatility and auto-regressive terms, we subsitute market prices with the total daily volume traded for the group of S&P 500 stocks in Equation 6. We run the regression for the logarithm of volumes and first differences of the logarithm, Table 15 reports the results. None of the coefficients for R-RISI are significant for the levels and the differences in volumes. Therefore, we cannot state any predictive capabilities of R-RISI on trading volumes.

Taken together with the Granger evidence, which establishes directional predictability from R-RISI to log SPX_VOL in a bivariate specification, and the level regressions above, which quantify the forecasting properties of retail sentiment net of volatility and autocorrelation, these results indicate that bullish online sentiment coincides with, but do not significantly precede, lower aggregate S&P 500 trading volume.

I Appendix: Structural break and robustness checks around the 2021 meme-stock episode

This appendix documents robustness checks for the main predictive regression in Section 6 around the January–February 2021 meme-stock episode. Two concerns motivate the exercise: (i) the 365-day rolling z-score window used to normalise R-RISI straddles the meme-stock spike during 2021–2022, so the scaling denominator itself may be affected; and (ii) the user composition on r/wallstreetbets changed materially after January 2021 (see Figure 2), potentially shifting the sentiment data-generating process.

Four checks are reported: Chow structural-break tests at three candidate dates; coefficient stability across three non-overlapping subsamples; a re-estimation with all 2021 observations excluded; and re-estimation of the main regression using R-RISI built from 180-day and 730-day rolling z-score windows as alternatives to the 365-day baseline.

Table 15: Effect of R-RISI on SPX volumes

	log(SPX_VOL)		Δ log(SPX_VOL)	
	γ_k	p-value	γ_k	p-value
Constant	7.5580	0.000	0.0237	0.359
$log(SPX\_VOL)_{t-1}$	0.2833	0.000		
$log(SPX\_VOL)_{t-2}$	0.1368	0.000		
$log(SPX\_VOL)_{t-3}$	0.0356	0.139		
$log(SPX\_VOL)_{t-4}$	0.0397	0.156		
$log(SPX\_VOL)_{t-5}$	0.1328	0.000		
$\Delta log(SPX\_VOL)_{t-1}$			-0.6509	0.000
$\Delta log(SPX\_VOL)_{t-2}$			-0.4614	0.000
$\Delta log(SPX\_VOL)_{t-3}$			-0.3807	0.000
$\Delta log(SPX\_VOL)_{t-4}$			-0.3000	0.000
$\Delta log(SPX\_VOL)_{t-5}$			-0.1228	0.000
$R-RISI_{t-1}$	-0.0139	0.147	-0.0360	0.259
$R-RISI_{t-2}$	-0.0099	0.324	-0.0228	0.503
$R-RISI_{t-3}$	0.0142	0.159	0.0460	0.188
$R-RISI_{t-4}$	0.0014	0.867	0.0066	0.820
$R-RISI_{t-5}$	-0.0077	0.466	-0.0201	0.561
VIX_{t-1}	0.0274	0.012	0.0643	0.105
VIX_{t-2}	0.0219	0.272	0.0679	0.301
VIX_{t-3}	-0.0095	0.754	-0.0821	0.403
VIX_{t-4}	0.0067	0.811	0.0148	0.876
VIX_{t-5}	-0.0188	0.491	-0.0849	0.380

Note: The table reports estimated coefficients for the model in Equation 6, replacing market prices with total trading volume as the dependent variable.

I.1 Specification recap

All tests use the main specification of Equation 6 of the SPX sample,

$$z_t^{\text{SPX}} = \alpha + \sum_{k=1}^{5} \beta_k z_{t-k}^{\text{SPX}} + \sum_{k=1}^{5} \gamma_k R\text{-}RISI_{t-k} + \sum_{k=1}^{5} \delta_k \text{VIX}_{t-k} + \phi' \text{DOW}_t + \varepsilon_t, \tag{9}$$

where z_t^{SPX} is the z-scored S&P 500 price, DOW_t contains Tuesday–Friday day-of-week dummies (Monday omitted), and standard errors are Newey–West (HAC) with 10 lags. The object of interest is the joint Wald F-test of $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0$.

I.2 Chow structural-break tests

Table 16 reports Chow F-tests of parameter constancy at two candidate break dates: January 27, 2021 (the GameStop peak) and June 1, 2021 (a post-meme-stock normalisation reference). The test statistic is the standard Chow (1960) test with $k = 20$ parameters (a constant plus five lags each of z^{SPX}, R-RISI, and VIX plus four DOW dummies).

Both 2021 break dates reject parameter constancy at the 1% level, consistent with the user-composition shift that followed the meme-stock episode.

I.3 Subsample stability

Table 17 re-estimates Equation (9) on four sub-samples: 2016–2020 (pre-meme-stock), 2021–2022 (during and immediately after the meme stock frenzy), 2023–2026 (post-normalisation, including the April 2025 tariff correction), and the whole post

Table 16: Chow structural-break tests on Equation (9)

Break date	F stat	df_1, df_2	p-value	N_{pre}	N_{post}	Signif.
2021-01-27 (GameStop peak)	1.921	(20, 2262)	0.008	1,122	1,180	***
2021-06-01 (post-meme reference)	1.953	(20, 2262)	0.007	1,196	1,106	***

Note: Chow F-test of the joint hypothesis that all coefficients of Equation (9) are constant across a single candidate break date. *** denotes rejection at the 1% level.

meme-stock period from 2021–2026 to cleanly seperate into period before and after a structural break.

Table 17: Subsample stability of the main price–sentiment regression

Sample	N	R^2	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\sum_k \hat{\gamma}_k$	Wald F	p-value
2016–2020	1,108	0.960	+0.0176	−0.0057	+0.0180	2.349	0.042
2021–2022	448	0.987	+0.0257	−0.0192	−0.0010	1.924	0.024
2023–2026	746	0.975	+0.0310	−0.0164	+0.0111	1.983	0.309
2021–2026	1,194	0.984	+0.0263	−0.0149	+0.0034	2.475	0.017
Full 2016–2026 (baseline)	2,302	0.974	+0.0204	−0.0087	+0.0094	4.214	0.001

Note: OLS estimates of Equation (9) with Newey–West HAC standard errors (10 lags). $\hat{\gamma}_1$ is the coefficient on R-RISI at lag 1; $\sum_k \hat{\gamma}_k$ is the sum of the five R-RISI lag coefficients. The Wald F tests the joint null $\gamma_1 = \cdots = \gamma_5 = 0$. Figure 30 plots the five lag coefficients and 95% HAC confidence intervals for each subsample.

The lag-1 R-RISI coefficient (see Figure 30) is positive in all four subsamples and in the full sample, and is larger in the two post-2020 regimes. The joint Wald F rejects at the 5% level in 2016–2020, 2021–2022, and 2021–2026 but not in 2023–2026 ($p = 0.309$). The 2023–2026 null reflects loss of power from the shorter sample rather than a sign change in the point estimates.

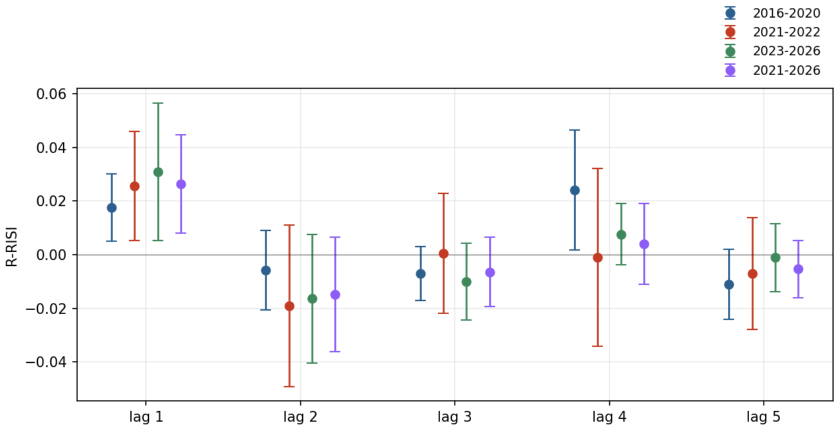
I.4 Excluding all of 2021

As a direct test of whether 2021 drives the headline result, Equation (9) is re-estimated with all 2021 calendar-year observations dropped. The remaining sample has $1,079$ observations. The lag-1 R-RISI coefficient is $\hat{\gamma}_1 = +0.0216$ ($p = 0.001$), and the joint Wald test gives 4.177 ($p = 0.001$). The headline price–sentiment result is therefore not driven by the meme-stock year.

I.5 Alternative rolling z-score windows

Table 18 rebuilds R-RISI with rolling-window z-scoring at 180-day and 730-day horizons, holding the paper's 365-day window as the baseline, and re-estimates Equation (9). The lag-1 R-RISI coefficient is statistically significant and positive in the baseline and the shorter 180-day specification, but looses validity in the longer 730-day window. The joint Wald test rejects at the 1% in the baseline and the short time window. The price-predictability result is therefore insensitive to the normalisation horizon for shorter periods, but looses significance when the rolling window goes to two years.

Figure 30: Subsample coefficient stability for R-RISI lags in Equation (9)



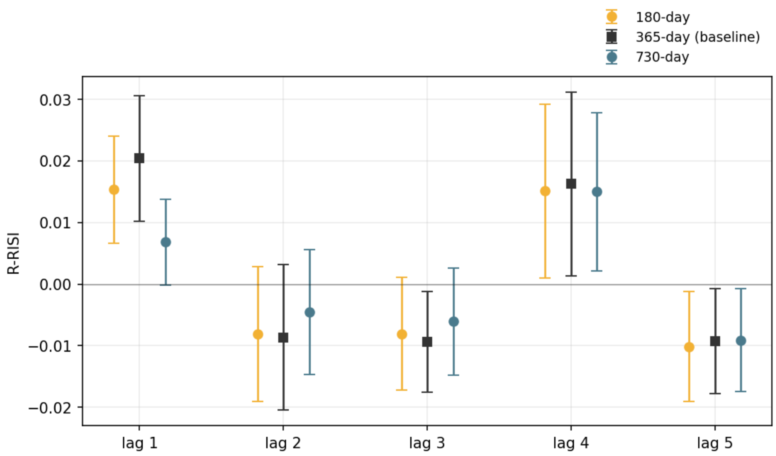
Note: R-RISI lag coefficients $\hat{\gamma}_1, \ldots, \hat{\gamma}_5$ with 95% HAC confidence intervals for the three non-overlapping subsamples (coloured markers) and the full-sample baseline (black dashed line with crosses). Error bars use $\pm 1.96 \cdot$ HAC-SE.

Table 18: Main regression under alternative R-RISI rolling z-score windows

Window	N	R^2	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\sum_k \hat{\gamma}_k$	Wald F	p-value
180-day	2,302	0.975	+0.0154	−0.0081	+0.0041	3.016	0.010
365-day (baseline)	2,302	0.976	+0.0204	−0.0086	+0.0094	4.214	0.001
730-day	2,302	0.976	+0.0069	−0.0046	+0.0021	1.736	0.123

Note: R-RISI is rebuilt by applying a rolling z-score to the weighted net sentiment series using 180-day and 730-day windows in addition to the paper's 365-day baseline.

Figure 31: R-RISI lag coefficients under alternative rolling z-score windows



Note: R-RISI lag coefficients $\hat{\gamma}_1, \ldots, \hat{\gamma}_5$ with 95% HAC confidence intervals for three rolling z-score windows.

Acknowledgements

We thank participants in the ECB seminar series of the Directorate-General for Macroprudential Policy and Financial Stability (DGMF) and the participants of the Barcelona School of Economics Summer Forum workshop ”Information Exchange and Trading in the Age of AI and Social Media: Theory and Experiments” for valuable feedback. We also thank Domenic Kellner and Luca Mingarelli for helpful comments.
The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank or the Eurosystem.

Fabian Wagner
European Central Bank, Frankfurt am Main, Germany; email: fabian.wagner@ecb.europa.eu

Julian Metzler
European Central Bank, Frankfurt am Main, Germany; University of Bristol, Bristol, United Kingdom;
email: julian.metzler@ecb.europa.eu

© European Central Bank, 2026

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

All rights reserved. Any reproduction, publication and reprint in the form of a different publication, whether printed or produced electronically, in whole or in part, is permitted only with the explicit written authorisation of the ECB or the authors.

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB’s website.

PDF ISBN 978-92-899-7990-0 ISSN 1725-2806 doi:10.2866/9838661 QB-01-26-204-EN-N