From Buzz to Bust: Sentiment Reversals and Return Predictability

Stockholm School of Economics, MSc in Finance Thesis
Luca Pitta (42696)

De-Biasing Venture Investments Through Alternative Data

Bocconi University, MSc in Finance Thesis
Luca Pitta (3093055)

Bocconi University Supervisor: Stefano Caselli, Associate Professor, Department of Finance, Bocconi University

Stockholm School of Economics Supervisor: Riccardo Sabbatucci, Associate Professor, Department of Finance, Stockholm School of Economics.

STOCKHOLM SCHOOL OF ECONOMICS
Department of Finance
4350 Master's Thesis in Finance
Fall 2025
© Luca Pitta, 2025

BOCCONI UNIVERSITY
Department of Finance
Master's Thesis in Finance
Fall 2024
© Luca Pitta, 2024

Abstract: In “From Buzz to Bust: Sentiment Reversals and Return Predictability,” I examine whether investor sentiment extracted from YouTube can help predict financial market returns. Using transcripts from financial YouTube videos, I filter for industry-relevant statements and use a large language model to assign sentiment scores, which I then aggregate into weekly industry sentiment indices. The results show that sentiment, when averaged over an 8-12 week window, is significantly and negatively related to next-week industry returns. This pattern is consistent with a contrarian, “sentiment-as-mispricing” interpretation: when the narrative gets too optimistic, returns tend to mean-revert. Beyond the findings, the thesis contributes a scalable way to turn YouTube content into a structured sentiment signal that can complement more traditional measures. In “De-Biasing Venture Investment Decisions,” I take the same broader idea, measuring narratives and crowd mood from alternative data, and bring it into the venture capital setting, where uncertainty and limited fundamentals make decision errors more likely. Framing the problem through behavioural finance, I focus on herding as the key bias and develop a model to quantify it through an original Herding Index built on social media sentiment. I also propose a refined way of measuring “success” in venture investments to better capture investment quality. Empirically, I find a significant negative relationship between herding and investment quality, particularly in periods of extremely high or low market returns, suggesting that holding a disciplined, contrarian view can pay off over the long run. Taken together, these two theses show how LLM-based analysis of online narratives can both uncover predictive signals in public markets and support more robust, less biased decision-making in private markets.

From Buzz to Bust: Sentiment Reversals and Return Predictability

1. Introduction **5**
2. Literature Review **7**
2.1 Market-wide timing using media tone 7
2.1.1 Cross-sectional stock selection from firm-level news 9
2.1.2 Blog and News Sentiment 10
2.2 Sentiment From Social Medias 11
2.2.1 Message Boards (Yahoo! Finance, Raging Bull) 12
2.2.2 Twitter/X Mood and Index Timing 12
2.2.3 Stock-level Social Media Sentiment (StockTwits, Twitter) 13
2.2.4 Reddit / WallStreetBets-Style Attention Strategies 14
2.3 Deep Learning & LLM-driven Sentiment Strategies 15
3. Sentiment and Market Data **19**
3.1 Sentiment Scores 19
3.1.1 Sentence Filtering 19
3.1.2 Sentiment Extraction 19
3.1.3 Sentiment Dataset 20
3.2 Market Returns 21
4. Results Analysis **22**
4.1 Regression Results 22
4.2 Interpretation 23
5. Conclusions **24**
Bibliography **26**
Appendix 1: Sentiment Data **29**
Appendix 2: Regressions Results **30**
Appendix 3: Use of AI **31**

e

1. Introduction

Investor sentiment usually refers to how optimistic or pessimistic investors are about future returns, beyond what one would expect from fundamentals alone. It is essentially the “mood” of the market: how investors feel about risk, about particular assets, or about the market as a whole, even when there is no obvious change in earnings, cash flows, or macroeconomic data. In the behavioural finance literature, sentiment is often used to explain mispricing, bubbles, crashes, and other situations where prices seem to move “too much” relative to news. In this sense, sentiment is not just a vague psychological concept, but something that can be treated as a systematic factor which we can attempt to measure and relate to asset prices.

Traditionally, investor sentiment has been measured in a few main ways. One important approach is survey-based, where investors are asked directly how bullish or bearish they are, and their answers are aggregated into indices. Typical examples include surveys of retail investors or professional forecasters. These measures are intuitive and straightforward to interpret, but they are usually low frequency, rely on relatively small and potentially unrepresentative samples, and may be affected by self-reporting bias.

A second group of measures is market-based, where sentiment is inferred indirectly from trading and issuance activity. Common proxies include trading volume, IPO activity and first-day returns, mutual fund flows, option-implied skewness, or the discount on closed-end funds. These indicators are available at higher frequency and are tightly linked to what investors actually do in markets. However, they mix sentiment with other forces such as liquidity needs, institutional constraints, and risk management considerations, which complicates the interpretation of the signal.

More recently, a large part of the literature has focused on text-based sentiment. Here, the idea is to extract sentiment from written or spoken language using natural language processing techniques. Existing work has relied on newspapers and financial newswires, earnings call transcripts, analyst reports, and, more recently, social media platforms such as Twitter, Reddit, or StockTwits. These sources allow us to move from simple quantities (like volume or flows) to something closer to tone, narratives, and themes. At the same time, they come with their own limitations. Traditional media primarily reflect the perspective of journalists and editors, who may be closer to institutional investors than to retail investors. Social media is more representative of retail participation, but the texts are often very short, informal, noisy, and strongly shaped by platform dynamics such as virality, bots and spam.

Investor sentiment measures derived from these different sources are widely used in both academic and applied work. In academia, they are used to explain cross-sectional patterns in returns, time-variation in expected returns, and the behaviour of variables such as volatility and trading volume. Sentiment has been linked to anomalies such as momentum, IPO mispricing, and the pricing of speculative or “hot” sectors. On the practitioner side, sentiment indicators are integrated into trading strategies, timing models, and risk management

frameworks. Asset managers may adjust their exposures when sentiment appears extremely optimistic or pessimistic, and policymakers increasingly monitor sentiment as part of their broader assessment of financial stability risks.

A central point in all of this is that the underlying data source matters. The information environment in financial markets has changed substantially over the past decade. One particularly important development is the rise of online video platforms, notably YouTube, as a major channel for financial content. Official channels like CNBC, Financial Times or The Wall Street Journal (among the others) can be considered as a good proxy of what investors are thinking, both retail and institutional ones.

From a research perspective, YouTube is a promising and still relatively underused data source. Compared to short-text platforms, YouTube videos are typically long-form and narrative-driven, which makes it possible to capture richer sentiment and more detailed narrative structures. In addition, metadata such as upload time, views, likes, and channel characteristics provide useful context about the potential reach and impact of a given message. Importantly, YouTube provides automatic transcripts for many videos, which can be collected and processed with standard text analysis tools. In practice, this means that a large body of spoken financial commentary can be turned into analysable text.

In this thesis, I do not aim to innovate on the trading strategy side. The objective is not to design or backtest a sophisticated strategy that consistently outperforms the market based on sentiment. Instead, the focus is on the source of sentiment: I construct and study a new sentiment measure derived from YouTube transcripts. The main question is whether sentiment extracted from YouTube - viewed as a proxy for the tone and narratives that retail investors are exposed to - contains information that is different from, or complementary to, the more traditional sentiment measures based on surveys, market variables, or other text-based sources.

More concretely, I assemble a dataset of YouTube videos related to financial markets and investing, obtain and clean their transcripts, and apply textual analysis techniques to quantify sentiment over time. I then compare this YouTube-based sentiment series to standard benchmarks from the literature, analysing its co-movement with existing measures, possible lead-lag relationships, and its behaviour around major market events. I also run a set of relatively simple empirical exercises to assess whether this new sentiment measure has any predictive content for returns or volatility.

Based on these findings, I then propose a trading strategy that can leverage this peculiar type of sentiment data, which allows it to significantly outperform its benchmark. This is only one of the many ways in which this data can be used, and some suggestions are then given for further developments.

2. Literature Review

In this thesis, two are the main topics being researched: sources of sentiment data, and trading strategies based on this data.

In the following, the thesis dives deep into each, unveiling what has been researched up until now and whether there exists any gaps that need to be filled.

2.1 Market-wide timing using media tone

A large literature shows that the tone of financial news co-moves with, and helps forecast, aggregate equity returns. The central idea is that media coverage both reflects and shapes investor sentiment. When news columns adopt an unusually pessimistic tone, some investors temporarily overreact, pushing prices away from fundamentals. As sentiment normalises, prices mean-revert. A systematic investor can therefore treat news tone as a market-wide timing signal, tilting index exposure in a contrarian way.

The seminal contribution is Tetlock (2007), who constructs a daily "pessimism index" from the Wall Street Journal's *Abreast of the Market* column over 1984–1999 using the Harvard General Inquirer dictionary[1]. He extracts a media "pessimism factor" as the first principal component of several negative word categories (e.g. Negative, Weak, Fail, Fall) and shows that high pessimism today robustly predicts lower Dow Jones Industrial Average returns tomorrow, even after controlling for lagged returns, volatility and calendar effects. The effect is economically meaningful: a one-standard-deviation increase in pessimism lowers next-day returns by roughly 8 basis points, larger than the unconditional mean of daily returns. Crucially, this effect is almost completely reversed over the following four trading days, so that the cumulative impact over a week is close to zero.

This pattern (short-horizon underperformance following pessimistic tone, followed by reversion) rules out simple "new information" stories in which negative media tone just reveals bad fundamentals. If the column merely conveyed news about lower future cash flows, prices should fall and stay low, not rebound within a week. Instead, the evidence is consistent with temporary downward price pressure stemming from low investor sentiment or higher risk aversion, as in noise-trader models. Tetlock also documents that extreme values of the pessimism index (both very high and very low) are associated with elevated trading volume, consistent with sentiment-driven rebalancing.

Garcia (2013) extends this approach both in time and methodology by analysing two financial columns in the *New York Times* from 1905 to 2005[2]. He measures news sentiment as the fraction of positive minus negative words and links it to indicators for NBER recessions. The key result is that predictability of stock returns from news sentiment is strongly concentrated in recessions: during downturns, negative tone predicts low contemporaneous returns and higher subsequent returns, whereas in expansions the predictive power of sentiment largely disappears. This state-dependence suggests that

[1] Tetlock, P.C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *Journal of Finance*, 62(3), 1139–1168.
[2] Garcia, D. (2013). "Sentiment During Recessions." *Journal of Finance*, 68(3), 1267–1300.

investors are more prone to sentiment-driven overreaction when macroeconomic uncertainty and risk aversion are elevated. It also implies that a conditional timing rule which leans against negative sentiment is particularly attractive “in bad times”, when insurance against downside risk is most valuable.

More recent work generalises these insights beyond US newspapers and simple dictionary counts. Shen et al. (2022) use the GDELT news database to construct several daily media sentiment indexes (tone, attention, polarity) for Chinese equity markets and show that these measures jointly have significant predictive power for domestic stock returns and conditional volatility, again with stronger effects around stressful periods[3]. Anese et al. (2023) apply modern NLP and long short-term memory (LSTM) networks to real-time financial news about the S&P 500 and document that public news sentiment impacts index returns and volatility at intraday horizons (within about 20 minutes), with dictionary-based labels outperforming return-based labels[4]. Together with machine-learning studies on large financial-news headline datasets, which show that sentiment extracted from headlines improves stock-return prediction and the profitability of associated trading rules[5], these papers suggest that the basic Tetlock-style patterns are robust to different markets, horizons and sentiment extraction techniques.

In practice, these strategies are often implemented as overlays on existing equity exposure rather than as stand-alone market-neutral trades. For example, a long-only fund might vary its net equity allocation between 80% and 120% of a benchmark index depending on a smoothed news sentiment score, while a risk-parity portfolio might use media tone as an additional risk-on/risk-off indicator alongside volatility and credit spreads. Recent work also combines media tone with other systemic indicators (for instance, enhancing financial stress indexes with GPT-based sentiment extracted from daily Bloomberg market summaries to improve Sharpe ratios and reduce drawdowns) suggesting that news-based timing can complement traditional macro and volatility signals[6].

Overall, market-wide timing strategies based on media tone rest on three robust empirical regularities: (i) aggregate news sentiment co-moves with investor sentiment and trading volume; (ii) unusually pessimistic tone tends to precede short-term underperformance of equity indexes, followed by mean reversion; and (iii) the strength and profitability of this pattern depend on the macro environment, being particularly pronounced in recessions and high-uncertainty regimes. These properties make media-tone timing a natural starting point for news-based equity strategies, before moving to finer-grained applications such as sector rotation or stock-level selection.

---

[3] Shen, S., Xia, L., Shuai, Y., & Gao, D. (2022). “Measuring News Media Sentiment Using Big Data for Chinese Stock Markets.” *Pacific-Basin Finance Journal*, 74, Article 101810.
[4] Anese, G., Corazza, M., Costola, M., & Pelizzon, L. (2023). “Impact of Public News Sentiment on Stock Market Index Return and Volatility.” *Computational Management Science*, 20(1), Article 20.
[5] Kamal, S., Sharma, S., Kumar, V., Alshazly, H., Hussein, H.S., & Martinetz, T. (2022). “Trading Stocks Based on Financial News Using Attention Mechanism.” *Mathematics*, 10(12), 2001.
[6] Lefort, B., Benhamou, E., Ohana, J.-J., Saltiel, D., Guez, B., & Jacquot, T. (2024). “Stress Index Strategy Enhanced with Financial News Sentiment Analysis for the Equity Markets.” *arXiv* preprint arXiv:2404.00012.

2.1.1 Cross-sectional stock selection from firm-level news

While early work focuses on market-wide sentiment, later studies shift the lens to firm-specific news and the cross-section of stock returns. The seminal contribution is Tetlock, Saar-Tsechansky and Macskassy (2008), who quantify language in Wall Street Journal and Dow Jones News Service articles for S&P 500 firms from 1980–2004 using the fraction of negative words from the Harvard General Inquirer dictionary. They show that firm-level negative tone forecasts both lower subsequent earnings and lower next-day abnormal returns, even after controlling for past returns, size, book-to-market and analyst forecasts. Positive words contain much weaker incremental information once negative tone is controlled for, so the predictive content of "news sentiment" at the firm level is heavily skewed towards bad news[7].

Tetlock et al. also explicitly translate their results into a long–short trading rule. Each day they classify a firm's DJNS stories as "negative" or "positive" based on whether the standardized fraction of negative words lies in the top or bottom quartile of its prior-year distribution, and then form an equal-weighted portfolio that is long firms with positive news and short firms with negative news, held for one day and rebalanced daily. In a frictionless setting this strategy earns a Fama–French / Carhart alpha of roughly 9-12 basis points per day (about 21% annualised), and the returns are largely orthogonal to standard size, value and momentum factors[8]. However, once realistic round-trip trading costs of around 10 basis points are imposed, the simple high-turnover implementation ceases to be profitable, highlighting that the signal is economically strong but costly to exploit at daily frequency.

Using a much larger proprietary Thomson Reuters news dataset, Heston and Sinha (2017) revisit firm-level news and emphasise asymmetric market reactions to good versus bad news. They find that positive stories tend to be incorporated into prices almost immediately, so their predictive power for future returns is short-lived (one to two trading days). By contrast, negative stories provoke a slower, more protracted adjustment, and weekly aggregated negative news predicts returns over several months, with much of the delayed reaction materialising around subsequent earnings announcements. This asymmetry implies that cross-sectional sentiment strategies often "work harder" on the short leg (betting against firms with persistently negative news flow) than on the long leg, where positive news is digested more efficiently[9].

Subsequent papers confirm and refine these patterns. Ahmad et al. (2016) build a large corpus of firm-specific media articles for U.S. blue-chip firms and show that media-expressed negative tone impacts firm-level returns episodically: in many periods, tone has little effect, but during specific episodes negative tone either induces sharp reversals or persistent drifts, depending on whether it mainly reflects transient sentiment or

---

[7] Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *Journal of Finance*, 63(3), 1437–1467.
[8] Carhart, M. M. (1997). "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52(1), 57–82; Fama, E. F., & French, K. R. (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33(1), 3–56.
[9] Heston, S. L., & Sinha, N. R. (2016). "News versus Sentiment: Predicting Stock Returns from News Stories." Finance and Economics Discussion Series 2016-048, Board of Governors of the Federal Reserve System.

value-relevant information[10]. Survey evidence in Kearney and Liu (2014) concludes more broadly that firm-level textual sentiment, whether from news, filings or conference calls, typically predicts the cross-section of stock returns at short horizons (days to a few weeks), with the strongest and most robust effects again linked to negative tone[11].

Across studies, such firm-level, news-based long–short portfolios generate economically meaningful pre-cost alphas (often several percent per year) when evaluated against multi-factor benchmarks, with much of the signal coming from the most negative-tone stocks. The main practical limitations are high turnover, sensitivity to transaction costs and capacity constraints, but the overall evidence supports the idea that firm-specific news sentiment is a distinct, short-horizon return predictor in the cross-section, complementary to traditional characteristics such as size, value and momentum.

2.1.2 Blog and News Sentiment

Zhang and Skiena (2010) push news-based trading into the realm of web-scale text, explicitly combining mainstream news with the long tail of blogs and microblogs to build trading agents[12]. Using the Lydia sentiment analysis system, they process roughly one terabyte of text from over 500 newspapers (“Dailies”), Spinn3r RSS blogs, LiveJournal and Twitter, matched to NYSE-listed firms over about five years (2005–2009). This setup allows them to ask not only whether firm-level sentiment predicts returns and volume, but also whether large, heterogeneous corpora of user-generated content add information beyond traditional financial news.

On the descriptive side, they show that media frequency is strongly linked to trading activity: normalised article counts and trading volume have correlations above 0.4 across media sources, with particularly strong relationships for sectors that were central to the subprime crisis (for example financials and housing-related firms). Sentiment polarity (a normalised measure of positive minus negative words per entity) is positively related to same-day stock returns, with little predictive power beyond one trading day for traditional news—consistent with rapid incorporation of newspaper information into prices. Blog sentiment behaves differently: polarity is more persistent and, for sources such as Twitter, there is some correlation with returns over the following two to three days, suggesting slower diffusion of information or opinions from the blogosphere into prices. Subjectivity (the share of sentiment-laden references) is also positively associated with trading volume, in line with earlier evidence that disagreement and controversy generate trades.

Building on these correlations, Zhang and Skiena design an explicit sentiment-based trading agent that operates as a market-neutral long–short portfolio. Each day, for each firm with sufficient coverage, Lydia produces sentiment measures (polarity, subjectivity and related ratios) from each corpus. Firms are then ranked by their sentiment score, and the strategy goes long an equal-dollar basket of the most positively discussed stocks and short an

[10] Ahmad, K., Han, J., Hutson, E., Kearney, C., & Liu, S. (2016). “Media-Expressed Negative Tone and Firm-Level Stock Returns.” *Journal of Corporate Finance*, 37, 152–172.
[11] Kearney, C., & Liu, S. (2014). “Textual Sentiment in Finance: A Survey of Methods and Models.” *International Review of Financial Analysis*, 33, 171–185.
[12] Zhang, W., & Skiena, S. (2010). “Trading Strategies to Exploit Blog and News Sentiment.” *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 375–378.

equal-dollar basket of the most negatively discussed stocks, producing an approximately dollar-neutral portfolio. The agent has several tunable parameters: (i) *n*, the number of long and short stocks (concentration versus diversification); (ii) *s*, the sentiment look-back window, from purely contemporaneous sentiment (*s* = 1) to multi-day averages; (iii) *h*, the holding period in days; and (iv) bounds on market capitalisation to restrict the investable universe. The strategy is implemented differently for news and blogs to respect timing: for newspapers they use same-day sentiment (most articles are spidered before the NYSE open), whereas for blogs they typically lag by one day because many posts appear after the open.

Backtests over 2005–2009 show that this “best-sentiment” strategy generates consistently positive returns with relatively low volatility, before transaction costs. Performance is highest for concentrated portfolios (small *n*), short sentiment windows (*s* close to 1) and short holding periods (one to two days), which is intuitive: very fresh, extreme sentiment provides the strongest signal, and the advantage dissipates quickly as the market absorbs the information. Both very small and very large firms appear more profitable than mid-cap names, either because small caps are more sensitive to sentiment shocks or because the sentiment of widely covered blue chips is measured more accurately. To validate that they are not simply picking up noise, the authors also test a “worst-sentiment” strategy (long the most negative, short the most positive) and a random-selection benchmark: the former systematically loses money, the latter fluctuates around zero, reinforcing that the sentiment sort has genuine directional content.

A key conceptual message of the paper is that “corpus size matters”: rather than relying solely on a small set of professional outlets such as DJNS or the *Wall Street Journal*, a broader aggregation of blogs and news yields a more reliable and powerful sentiment signal. In other words, the long tail of web commentary contains incremental information (or at least incremental sentiment) that is not fully captured by mainstream financial media, and this can be harvested by automated long / short equity strategies. At the same time, the reported results ignore transaction costs and potential implementation frictions, so the net profitability for real-money investors is uncertain; but Zhang and Skiena provide an important proof of concept that large-scale blog and news sentiment can be translated into systematic, market-neutral stock selection rules.

2.2 Sentiment From Social Medias

Compared with traditional financial news, online discussions and social media provide much cheaper, higher-frequency and more retail-dominated signals about sentiment and attention. Message boards, microblogs and Reddit forums are populated mainly by individual investors and speculators rather than institutional analysts, so they are especially relevant for very short horizons and for stocks whose ownership base is dominated by retail investors. Across this literature a few recurring themes emerge. First, measures of “bullishness” or net optimism on these platforms have statistically significant but often economically modest power to forecast returns at horizons of one or a few days. Second, measures of message volume and disagreement are much more powerful predictors of volatility and trading volume. Third, even when raw predictive power is present, net profitability is highly sensitive to transaction costs, execution latency and limits to arbitrage.

2.2.1 Message Boards (Yahoo! Finance, Raging Bull)

Antweiler & Frank (2004) provide one of the earliest systematic analyses of online stock message boards. They study more than 1.5 million posts from Yahoo! Finance and Raging Bull for a sample of large “old economy” and “new economy” firms, including major constituents of the Dow Jones Industrial Average and the Dow Jones Internet Index[13]. Using a simple text classification scheme, each message is labelled as a buy, sell or hold recommendation, which allows them to construct daily measures of message “bullishness” (net buy minus sell messages), overall posting volume and a proxy for disagreement based on the dispersion of opinions across contributors.

Their main results can be summarised as follows. First, message-board bullishness contains statistically significant predictive power for next-day stock returns: days with more bullish messages tend to be followed by slightly higher abnormal returns, and days with more bearish messages are followed by slightly lower returns. However, the magnitude of these effects is small relative to daily volatility, so the implied Sharpe ratios are limited once realistic frictions are introduced. Second, posting volume and disagreement are strongly associated with next-day trading volume and return volatility. Periods characterised by many posts and highly divergent views foreshadow more turbulent trading days, consistent with models in which disagreement and speculative trading generate volume and volatility. Third, message boards appear to reflect the activities of short-horizon traders rather than insiders with superior fundamental information: most of the predictive content is exhausted within a day, and there is little evidence of longer-horizon drift associated with message sentiment.

From a strategy-design perspective, Antweiler & Frank motivate two basic uses of message-board data. One is short-horizon directional trading in individual names: a trader could increase long exposure after spikes in bullishness and become more cautious or even short after bursts of bearishness, typically with a one-day holding period. The other is volatility trading: message volume and disagreement can be used to forecast volatility and to time positions in options or other volatility-sensitive instruments, increasing exposure when message-board activity signals an impending volatility spike. In practice, once bid–ask spreads, commissions, short-sale constraints and capacity limits are taken into account, the economic significance of these signals is modest. Nevertheless, the paper is foundational in establishing the idea that “forums are signals”: even relatively unstructured online chatter contains information that is reflected in subsequent trading and prices, rather than being pure noise.

2.2.2 Twitter/X Mood and Index Timing

Bollen, Mao & Zeng (2011) extend the analysis from firm-specific forums to aggregate public mood as expressed on Twitter[14]. Using millions of tweets from 2008, they construct daily mood indices with text-analysis tools such as OpinionFinder (capturing positive versus negative sentiment) and GPOMS, which maps words into six psychological dimensions (for example Calm, Happy, Sure). These time series are then related to daily changes in the Dow

[13] Antweiler, W., & Frank, M. Z. (2004). “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” *Journal of Finance*, 59(3), 1259–1294.
[14] Bollen, J., Mao, H., & Zeng, X. (2011). “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science*, 2(1), 1–8.

Jones Industrial Average (DJIA). The key empirical result is that certain mood components (notably the Calm dimension) Granger-cause DJIA moves at horizons of one to six days. When lagged mood indices are included in nonlinear forecasting models (for instance, a self-organising fuzzy neural network), out-of-sample classification accuracy for DJIA up/down moves improves relative to simple autoregressive benchmarks, and forecast errors are reduced.

This naturally suggests Twitter-based index timing strategies. In a typical implementation, the investor would obtain a firehose feed of tweets, compute daily mood indices, and train a simple classifier such as logistic regression or a neural network to predict whether the DJIA (or a close proxy such as SPY) will rise or fall the next day. The trading rule would then be to go long DJIA futures or ETFs when the model predicts an up move and to go short or remain flat otherwise, possibly scaling exposure by model confidence or by the extremeness of the mood signal. Because Twitter data are available in almost real time, such a strategy can, in principle, be updated and re-estimated frequently.

Subsequent work both refines and challenges these findings. On the refinement side, later papers use finance-specific lexicons, more sophisticated sentiment classifiers and transformer-based language models to derive mood indices that are more directly related to financial concepts. These studies often find that Twitter sentiment has some predictive content for index returns and, perhaps more robustly, for volatility and risk measures, but the effect sizes are usually smaller than those reported in early work[15].

2.2.3 Stock-level Social Media Sentiment (StockTwits, Twitter)

A later strand of the literature shifts from aggregate mood to stock-specific messages, exploiting the fact that Twitter and StockTwits posts often include ticker identifiers (for example, “$TSLA”). This makes it possible to construct firm-level sentiment and attention measures analogous to those used in the news-based literature, but at much higher frequency and with a user base that is heavily tilted towards retail traders. Studies such as Sprenger et al. (2014) show that Twitter bullishness for S&P 100 firms—defined as the fraction of positive versus negative tweets tagged with a given ticker—is positively correlated with same-day abnormal returns and trading volume, and that spikes in message volume and disagreement are associated with elevated volatility and turnover[16]. The evidence for predictive power beyond the same day is more mixed: for large, liquid blue chips, social-media sentiment often co-moves with prices but does not clearly lead them, while for smaller, more speculative stocks there is some evidence that sentiment shocks precede return and volume movements over horizons of one trading day or less.

StockTwits, which is explicitly designed as a financial microblogging platform, has become a particularly popular data source for this type of analysis. Early work using coarse or dictionary-based sentiment measures typically found only weak unconditional predictability of returns. More recent studies that employ machine-learning classifiers to label posts as

---

[15] Nguyen, T. H., Shirai, K., & Velcin, J. (2015). “Sentiment Analysis on Social Media for Stock Movement Prediction.” In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1357–1362.
[16] Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). “Tweets and Trades: The Information Content of Stock Microblogs.” *European Financial Management*, 20(5), 926–957.

bullish, bearish or neutral, and then aggregate these labels to stock–day polarity measures, document stronger links between sentiment and abnormal returns around salient events such as earnings announcements or sudden peaks in message activity[17] [18]. A recurring finding is that the combination of a sentiment spike and a sudden increase in message volume is especially informative: when many users suddenly start posting strongly bullish (bearish) messages about a stock, that episode tends to coincide with or slightly precede large positive (negative) abnormal returns, followed by at least partial reversal as attention fades.

These empirical patterns underpin two broad types of strategy. The first is intraday momentum around sentiment bursts: an investor monitors cashtag streams for abrupt jumps in both sentiment and message volume, and then trades in the direction of the sentiment (buying on positive bursts, shorting on negative bursts) while closing positions quickly, typically within minutes or hours. The second is cross-sectional long / short stock selection based on daily aggregated sentiment. In that case the investor computes abnormal sentiment for each stock relative to its own recent history or to the cross-section, ranks the universe by this metric, and goes long the most positively discussed names and short the most negatively discussed ones, often imposing size and sector neutrality and holding positions for a few days. In practice, both approaches face familiar challenges around execution costs, short-sale constraints and capacity, but the pre-cost alphas reported in the literature are often economically meaningful, especially in retail-dominated small-cap segments.

Recent work experiments with more complex machine-learning architectures, including graph neural networks (GNNs), to model the network structure linking users, stocks and messages. In such models, nodes represent users or stocks and edges reflect relationships such as co-mentions, replies, follower links or shared sector membership. By propagating sentiment and attention shocks through this graph, GNN-based predictors can exploit information not only in the sentiment of individual messages but also in how information spreads across communities and across related stocks. Empirical results suggest that incorporating these network effects improves the accuracy of return forecasts from social-media data and enhances the performance of sentiment-based trading strategies, particularly during turbulent market conditions when retail activity is elevated[19] [20].

2.2.4 Reddit / WallStreetBets-Style Attention Strategies

Reddit's r/WallStreetBets (WSB) provides perhaps the most vivid recent example of social-media-driven market dynamics. WSB is an online forum where largely anonymous retail traders discuss high-risk, high-leverage trades, often in highly speculative stocks and short-dated options. Papers such as Warkulat & co-authors (2024) study how WSB attention

[17] Divernois, M.-A., & Filipović, D. (2024). "StockTwits Classified Sentiment and Stock Returns." *Digital Finance*, 6, 249–281.
[18] Gabrovšek, P., Aleksovski, D., Mozetič, I., & Grčar, M. (2017). "Twitter Sentiment Around the Earnings Announcement Events." *PLOS ONE*, 12(2), e0173151.
[19] Si, J., Mukherjee, A., Liu, B., Pan, S. J., Li, Q., & Li, H. (2014). "Exploiting Social Relations and Sentiment for Stock Prediction." In *Proceedings of EMNLP 2014*, 1139–1145.
[20] Zhang, P., Harris, R. D. F., & Zheng, J. (2024). "GNN-Based Social Media Sentiment Analysis for Stock Market Forecasting and Trading." SSRN Working Paper No. 5051165.

affects retail portfolios and welfare[21]. Using brokerage data, they show that when a stock receives increased attention on WSB, retail investors tend to tilt their portfolios towards that stock and towards riskier positions more generally: they allocate larger weights to WSB-mentioned names, trade more frequently and increase their exposure to idiosyncratic risk. Ex post, however, these WSB-induced trades are associated with lower average returns than a suitably matched benchmark, implying that the attention episodes reduce investor welfare on average rather than enhancing it.

From these episodes, several strategy templates can be extracted, although all of them are inherently high risk. A first category is pump-chasing or meme-following strategies. Here the trader continuously monitors WSB for sharp increases in the number of mentions and for strongly positive language around specific tickers. The idea is to buy very early in the attention spike and to sell into the subsequent wave of retail buying and increased liquidity. Because prices move extremely fast during such episodes, the profitability of this approach is highly path-dependent: being even a few hours late can turn an ostensibly attractive trade into a large loss. A second category consists of attention-reversion trades. In this case the investor waits until social-media attention and prices have clearly peaked, then shorts or underweights the previously targeted names, betting on a gradual return towards more fundamental valuations as the hype cycle fades. In practice, implementation is complicated by short-sale constraints, high borrow costs and the risk of further short squeezes.

A third application uses WSB activity as a volatility signal in options and volatility markets. Because attention spikes are associated with very large realised volatility and with extreme moves in implied volatility, an investor might choose to buy volatility (for example via call options, straddles or variance swaps) ahead of anticipated hype, or to sell volatility once implied volatilities have become extraordinarily elevated relative to historical norms and to fundamentals. Again, execution is sensitive to timing: implied volatility can remain high for extended periods during a frenzy, so selling too early can be costly, while buying too late will fail to capture the initial repricing.

Taken together, the Reddit/WSB literature shows that social-media attention can have powerful short-term effects on prices, volumes and volatility, particularly in heavily shorted or otherwise constrained stocks. However, it also underlines that exploiting these episodes is difficult and risky. For traditional institutional investors, WSB-type signals may be more useful as risk indicators or constraints (for example, as triggers to reduce risk or tighten risk limits in vulnerable names) than as a stable source of alpha.

2.3 Deep Learning & LLM-driven Sentiment Strategies

The most recent wave of work on news- and sentiment-based trading moves beyond dictionary counts and simple bag-of-words models towards deep neural networks and, more recently, large language models (LLMs). The common aim is to capture richer linguistic context, non-linear interactions and cross-asset structure that earlier methods largely miss. In empirical finance, the applications still revolve around familiar strategy archetypes (market timing, cross-sectional long / short, event-driven trading), but the construction of the underlying sentiment or “text factor” has become far more sophisticated.

[21] Warkulat, S. (2024). “Social Media Attention and Retail Investor Behavior.” *International Review of Financial Analysis*, 96, 102640.

A first strand focuses on deep neural sentiment models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformer encoders trained on financial text. FinBERT, a BERT-style transformer pre-trained on generic corpora and then fine-tuned on large collections of financial news and disclosures, is the canonical example. It delivers systematically higher accuracy than generic BERT and traditional machine-learning classifiers on benchmark datasets such as Financial PhraseBank and FiQA, and thus produces more reliable positive/negative/neutral labels for financial statements[22]. In applied work, FinBERT-type models are either used as drop-in replacements for dictionary scores (producing daily firm or index-level sentiment indices that feed into otherwise standard timing or cross-sectional strategies) or embedded as the text branch of more complex prediction architectures.

Deep "end-to-end" frameworks typically integrate textual features with price and other numerical data in a single model. Chen (2024), for instance, proposes a multimodal deep-fusion architecture in which one branch processes historical prices and technical indicators through an LSTM network, while a parallel FinBERT branch extracts sentiment features from daily news headlines; a fusion layer then combines these representations to predict next-day stock trends[23]. The combined model substantially outperforms purely price-based networks and naive concatenations of price data with bag-of-words features, both in terms of directional accuracy and in simulated trading performance. Similar multimodal designs now appear in factor-model settings, where deep networks are trained to extract latent factors jointly from prices, news and sometimes social-media sentiment, and the resulting expected-return and risk estimates are plugged into portfolio-optimisation routines.

The cutting edge, however, is explicitly LLM-based sentiment and trading. Here the goal is no longer just to classify short snippets of text as positive or negative, but to exploit the full contextual and generative capabilities of LLMs to measure financial sentiment, interpret complex narratives and generate trading signals. Kirtac and Germano's series of papers is representative. In "Sentiment Trading with Large Language Models" and "Enhanced Financial Sentiment Analysis and Trading Strategy Development Using Large Language Models", they analyse a corpus of 965,375 U.S. financial news articles from 2010–2023 and compare multiple sentiment engines: Loughran–McDonald dictionary scores, BERT and FinBERT classifiers, and GPT-style models such as OPT[24][25]. They show that LLM-based sentiment, and in particular OPT, attains higher accuracy in predicting the sign of subsequent stock returns and materially increases the explanatory power of sentiment factors in panel regressions. A simple self-financing strategy that each day goes long the quintile of stocks with the highest OPT-based sentiment and short the quintile with the lowest

[22] Araci, D. (2019). *"FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*." arXiv preprint 1908.10063.
[23] Chen, P., Boukouvalas, Z., & Corizzo, R. (2024). "A Deep Fusion Model for Stock Market Prediction with News Headlines and Time Series Data." *Neural Computing and Applications*, 36, 21229–21271.
[24] Kirtac, K., & Germano, G. (2024). "Sentiment Trading with Large Language Models." *Finance Research Letters*, 62, 105227.
[25] Kirtac, K., & Germano, G. (2024). "Enhanced Financial Sentiment Analysis and Trading Strategy Development Using Large Language Models." In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA)*, 1–10.

achieves a Sharpe ratio of roughly 3.0 over the sample (after a realistic allowance for transaction costs), compared with about 1.2 for an analogous strategy using dictionary sentiment, suggesting that better text understanding directly translates into stronger trading performance.

Beyond static classification, LLM-based sentiment is increasingly embedded in more complex trading systems. One common pattern is to use LLMs as high-quality labellers: an LLM is prompted or fine-tuned to annotate large unlabeled news corpora with sentiment scores or event tags, and lighter supervised models are then trained on this expanded dataset for deployment in production. Another line of work integrates LLM-derived sentiment directly into reinforcement-learning frameworks for portfolio optimisation. In these setups, the agent observes both market data and LLM-based sentiment features and learns trading policies that maximise risk-adjusted returns. Recent work on sentiment-augmented Proximal Policy Optimisation (SAPPO), for example, injects daily asset-level sentiment extracted from financial news into both the state representation and the advantage function, and reports systematically higher Sharpe ratios and lower drawdowns compared with price-only PPO baselines[26].

Central banks and policy institutions have also begun to adopt LLM-based sentiment tools. A Swiss National Bank working paper on FX sentiment analysis fine-tunes state-of-the-art open-source LLMs on labelled foreign-exchange news and compares their performance with traditional sentiment models and alternative machine-learning approaches. The fine-tuned LLMs not only achieve higher classification accuracy for the tone of FX articles, but also support trading strategies in G10 currency pairs that outperform those based on older sentiment measures in terms of both forecast metrics and portfolio returns[27].

Across these deep-learning and LLM-based approaches, the strategy patterns themselves remain familiar. On the time-series side, researchers continue to build index-level sentiment or text factors and use them for market timing: when the model infers an unusually negative tone in aggregate news or social media, the strategy cuts equity or FX beta; when sentiment improves or mean-reverts, risk is scaled back up. On the cross-sectional side, firm-level signals derived from FinBERT-type classifiers, LLM-based factors or embedding models are used to rank stocks each day, going long those with the most favourable news flow and short those with the worst, typically in beta- and sector-neutral portfolios. Event-driven variants focus on intraday windows around earnings announcements, macro releases or major idiosyncratic news, where the speed and accuracy of LLM sentiment extraction matter most. The main difference relative to earlier generations of the literature is therefore not the trading logic but the measurement technology. Deep neural networks and LLMs provide more accurate and nuanced estimates of the information and sentiment embedded in financial text, and can exploit relational and multimodal structure that bag-of-words or dictionary approaches ignore. In backtests, this typically appears as higher predictive accuracy, stronger factor Sharpe ratios and better performance of sentiment-based strategies—but at the cost of greater model complexity, heavier computation and more acute risks of overfitting

---

[26] Kirtac, K. (2025). “Sentiment-Augmented Reinforcement Learning for Portfolio Optimisation with Large Language Models (SAPPO).” *Leveraging LLM-Based Sentiment Analysis for Portfolio Optimization*, REALM 2025 / OpenReview.
[27] Ballinari, D., & Maly, J. (2025). *“FX Sentiment Analysis with Large Language Models*.” Swiss National Bank Working Paper No. 2025-11.

and structural instability. In practice, these models are often used in combination with simpler, more interpretable sentiment measures, with LLM-based signals treated as an additional layer of information rather than as a wholesale replacement for the tools developed in the earlier literature.

3. Sentiment and Market Data

Overall, we collected transcripts from more than 50,000 YouTube videos, resulting in a dataset of over 10 million sentences. We then processed the raw data to extract sentiment toward the different topics mentioned in these sentences.

3.1 Sentiment Scores

To extract sentiment scores from the raw transcripts, we first filtered the dataset to retain only sentences relevant to our research scope. We then computed a sentiment score for each retained sentence.

To ensure domain relevance, we restricted the sample to YouTube channels focused on financial news such as CNBC, the Financial Times, and Bloomberg and collected data from the early 2010s up until now.

3.1.1 Sentence Filtering

As explained later, the analysis is based on the sentiment around certain industries. To select the relevant sentences, we used a set of keywords for each of the industries as shown below. For example, find below the ones used for the Technology industry.

```
KEYWORDS = ["Tech", "tech", "Technology", "technology", "Software",
"software", "Hardware", "hardware", "AI", "Artificial Intelligence",
"artificial intelligence"]
```

The industries were selected primarily based on how reliably sentiment could be extracted for each of them. In practice, this meant choosing industries for which clear, unambiguous keywords could be identified: keywords that directly refer to the industry in question.

A further filtering step was applied during scoring: as explained in the next subsection, all observations assigned a score of “3” were excluded from the final sample.

3.1.2 Sentiment Extraction

In this thesis, I use large language models (LLMs) to extract sentiment because, as discussed in Section 2, they are particularly well-suited to this task compared to traditional NLP approaches, mainly due to their stronger ability to capture context.

Sentiment scores were computed using OpenAI’s GPT-4o-mini model (training data up to July 2024). To avoid potential contamination from the model’s training horizon, the dataset was restricted so that all sentiment observations occur after that date[28]. This ensures that the model is not evaluating content it may have already encountered during training.

[28] For a full list of models, please refer to OpenAI’s list here: https://platform.openai.com/docs/pricing

We used the following system prompt:

"You are my private banker and need to advise me on the market movements around the `{Industry}` *industry stocks. I will give you some sentences, and all you need to do is to tell, based on that sentence, if you would go long on the market or short. You need to do so by giving me a score that ranges from 1 to 5, where 1 means definitely short, 5 means definitely long, and 3 means neutral. Any score in between these is an intermediate one."*

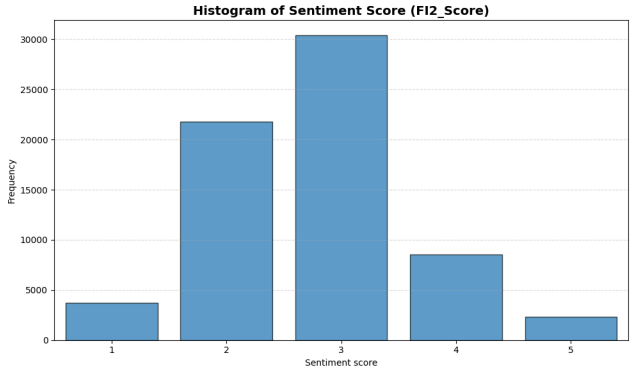
After collecting the sentiment scores, we further cleaned the dataset by removing all observations assigned a score of "3". These observations do not contribute meaningful information to the sentiment signal and mostly represent noise—typically sentences that are unrelated to stock returns.

This filtering step would be difficult to implement reliably with simpler sentiment models (e.g., BERT), which tend to rely more heavily on surface-level cues. In contrast, an LLM can use context to distinguish genuinely finance-relevant statements from irrelevant content. As a result, we can substantially reduce the sample to the observations that are most informative for the analysis.

3.1.3 Sentiment Dataset

Below are summary statistics for the collected data. As shown, the scores are concentrated around 3, an outcome that is expected, since a score of 3 corresponds to observations deemed irrelevant. These are sentences that the Large Language Model (LLM) classifies as having no meaningful impact on financial markets, and they are therefore excluded from the final sample.

A full analysis of the dataset is provided in Appendix 1. The table below reports summary statistics for the Financial Industry subsample only.



Summary statistics of sentiment score (FI2_Score)

	Value
Count	66667
Mean	2.76
Std dev	0.87
Skew	0.29
Kurtosis	0.17

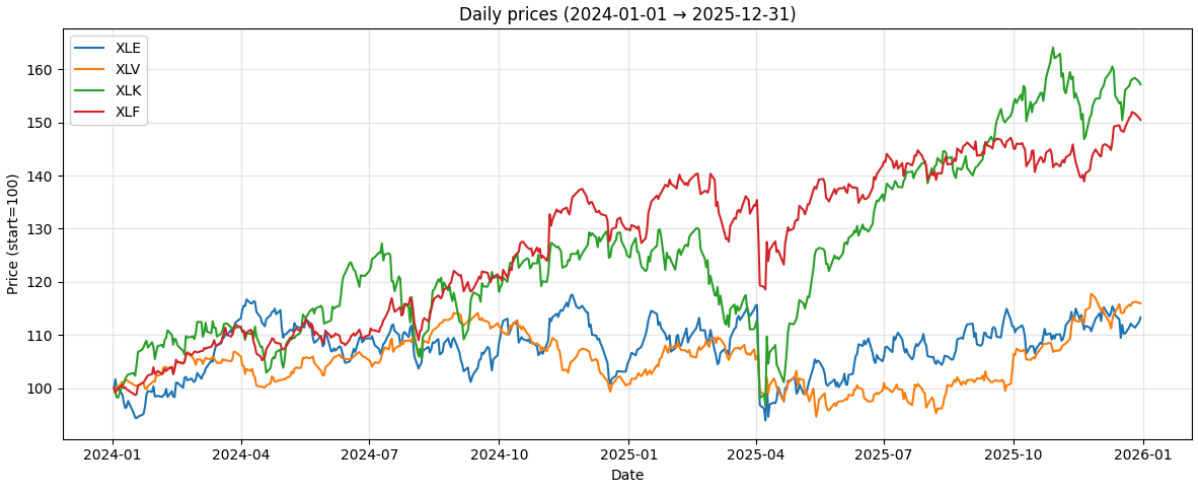
Some preliminary observations are worth noting. As shown later, the correlation between sentiment scores and market returns differs across industries; these relationships will be formally estimated in the empirical section. Even from an initial inspection, however, it is clear that some industries are "noisier" than others, as indicated by a larger share of observations scored as "3", i.e., sentences that are entirely irrelevant for the purposes of this analysis.

The scores also provide an initial indication of investor appetite over the sample period. Technology stands out in particular, with an average score above 3 and a clearly negatively skewed distribution, suggesting a stronger tilt toward positive sentiment.

For the main analysis, sentiment is aggregated at a weekly frequency for two reasons. First, the literature reviewed in Section 2 commonly adopts weekly windows as a good compromise for capturing sentiment dynamics: daily measures can be excessively noisy, while monthly aggregation may smooth away meaningful variation. Second, because the sample is truncated at July 2024, using weekly aggregation ensures a sufficient number of observations to run statistically meaningful regressions.

3.2 Market Returns

As dependent variable, we used market returns[29] for all the industries analysed by leveraging the S&P Industry Indexes[30]. These indices are widely used in the literature and, thanks to their high liquidity, they also provide a practical benchmark that could be traded on in future applications.



As we can see, Technology (XLK) is the best performing industry, in line with the sentiment data collected and shown in Appendix 1.

[29] Yahoo Finance, 2026
[30] Source: S&P 2026 (https://www.spglobal.com/spdji/en/index-family/equity/us-equity/sp-sectors/#overview)

4. Results Analysis

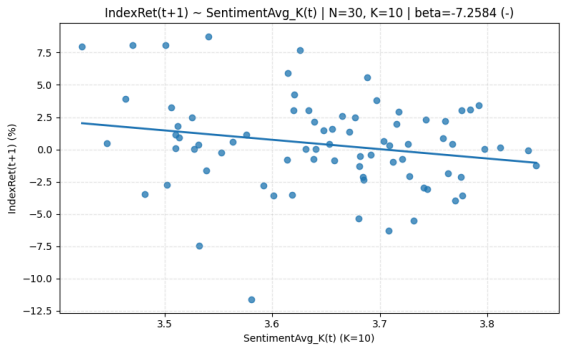
To study the link between sentiment and returns, we estimate a set of linear regressions at multiple time horizons (daily, weekly, and monthly). Across specifications, weekly average sentiment delivers the strongest association with industry returns.

This pattern is intuitive and consistent with prior evidence. Daily sentiment can be sparse and noisy, and, when it does matter, its predictive content tends to be short-lived (often concentrated in the next 1–2 trading days). By contrast, aggregating sentiment weekly helps reduce high-frequency noise and can uncover more persistent return predictability[31]. Finally, monthly aggregation may be too smooth to capture meaningful week-to-week swings in sentiment, since lower-frequency indices are typically designed to smooth idiosyncratic variation and reflect slower-moving components[32].

4.1 Regression Results

We ran separate regressions for each index and found a statistically significant negative relationship between the weekly average sentiment and the following week's return. In particular, our analysis indicates that sentiment averaged over an 8-12 week window (i.e., roughly 2-3 months) is the strongest predictor and therefore the most suitable independent variable.

We report the results for the Financials index here below; the full set of results is provided in Appendix 2.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.044
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	7.078
Date:	Tue, 06 Jan 2026	Prob (F-statistic):	0.00954
Time:	22:08:02	Log-Likelihood:	-205.24
No. Observations:	77	AIC:	414.5
Df Residuals:	75	BIC:	419.2
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	26.8763	9.875	2.722	0.008	7.204	46.549
x1	-7.2584	2.728	-2.661	0.010	-12.693	-1.824

Omnibus:	7.641	Durbin-Watson:	2.368
Prob(Omnibus):	0.022	Jarque-Bera (JB):	8.962
Skew:	-0.457	Prob(JB):	0.0113
Kurtosis:	4.400	Cond. No.	140.

The OLS estimates indicate a statistically significant negative association between returns and sentiment. The coefficient indicates a strong negative relationship: for each unit change in sentiment, a change of 7.26% follows holding other factors constant. Explanatory power is extremely limited (<5%), and combining this with other control variables could yield more significant results and would definitely be the first step to take for future research. The intercept is positive and significant, although sentiment (in our model) could never reach the

[31] Heston, S. L., & Sinha, N. R. (2016). News versus Sentiment: Predicting Stock Returns from News Stories (FEDS Working Paper 2016-048).
[32] Liu, Q., & Son, H. (2024). Methods for aggregating investor sentiment from social media. Humanities and Social Sciences Communications, 11, Article 925.

value of 0. Residual diagnostics (Omnibus and Jarque–Bera tests) suggest departures from normality, with moderate negative skewness and elevated kurtosis, while the Durbin–Watson statistic (2.37) does not indicate problematic positive serial correlation. Overall, the results support a robust negative relationship, but they also point to the need for richer model specifications to improve fit and to account for remaining unexplained variation. The same results can be inferred from all the other regressions.

4.2 Interpretation

Across all analysed industries, high average sentiment sustained over the previous 10 weeks predicts lower returns in the following week. The consistency of this relationship suggests that prolonged optimism builds expectations that become difficult to satisfy; when reality fails to keep up, prices partially reverse. Symmetrically, extended pessimism appears to set expectations too low, leading to positive reversals.

This pattern is coherent with the “sentiment-as-mispricing” view: when optimism stays elevated for a while, prices can drift above fundamentals, and the correction shows up as lower subsequent returns (a contrarian predictability). Baker and Wurgler explicitly document that high sentiment is followed by lower subsequent market returns, and (in the cross-section) that high sentiment periods are associated with lower subsequent returns[33]. Media-based sentiment measures also display a “pressure then reversion” pattern: Tetlock finds that pessimistic media content forecasts downward pressure followed by reversion, which is the same correction mechanism that implies optimistic sentiment can be followed by weaker subsequent returns[34].

Importantly, recent work using large-scale investor social media arrives at the same reversal-style implication. Cookson, Lu, Mullins, and Niessner build market-wide sentiment indices from millions of posts across major investor social platforms and show that sentiment is extrapolative and exhibits a strong reversal, consistent with elevated sentiment predicting weaker returns ahead[35]. At the security level, Watanabe uses LLM-based sentiment on StockTwits for OTC equities and finds that prices rise with positive sentiment and reverse afterwards, again matching the negative lagged sentiment / next-week return relation when sentiment is persistently high[36]. Evidence around discrete information events points in the same direction: Lopez Avila, Martineau, and Mondria show that pre-earnings social media attention is overly optimistic, creates buying pressure, and is followed by a return reversal when markets correct the distortion[37].

---

[33] Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129–152.
[34] 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.
[35] Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2025). Market signals from social media. SSRN Working Paper.
[36] Watanabe, M. (2025). Social media sentiment and over-the-counter equity returns. Working paper (June 30, 2025).
[37] Lopez Avila, E., Martineau, C., & Mondria, J. (2024). Social media and the distortion of price revelation. SSRN Working Paper (May 16, 2024).

5. Conclusions

This thesis aimed to assess whether long-form financial commentary from YouTube can be transformed into a useful and distinct investor-sentiment signal, and whether that signal contains information about subsequent market returns beyond what is typically captured by traditional survey-, market-, and text-based sentiment measures. To do so, the thesis constructed a novel sentiment dataset from YouTube video transcripts published by large financial-news channels, extracted sentence-level sentiment using a large language model, and aggregated the resulting signal at the industry level to study its relationship with future returns.

We find that the YouTube-based sentiment measure is systematically related to subsequent industry returns, with the strongest association emerging at weekly horizons. In particular, sentiment averaged over an 8-12 week window is most strongly (and robustly across specifications) negatively related to next-week returns. The overall pattern is consistent with a contrarian interpretation of sentiment: extended periods of elevated optimism tend to coincide with (or contribute to) over-extrapolation and gradual mispricing, which is later corrected through weaker subsequent returns; conversely, sustained pessimism is followed by relatively stronger returns. While the explanatory power of single-variable regressions remains modest, as is typical in return-prediction settings, the empirical results suggest that YouTube transcripts capture a sentiment component that is economically meaningful and potentially complementary to more established proxies. The main contribution of the thesis is therefore methodological and empirical: it demonstrates (i) that YouTube transcripts can be scaled into a structured sentiment signal using modern LLMs, and (ii) that this alternative sentiment source contains predictive content at medium horizons for industry-level returns, supporting the broader idea that the “information environment” faced by investors now extends well beyond newspapers, filings, and short-text social media.

At the same time, several limitations should be acknowledged. The sample is necessarily constrained by the post-July 2024 design choice to avoid overlap with the model’s training horizon, which limits the number of observations and the range of market regimes covered. The approach also depends on transcript quality, channel selection, and the assumptions embedded in LLM-based scoring (prompt design, model stability, and potential classification drift). Moreover, regressions should be better tested with control variables and other new elements that can prove their robustness. Finally, any trading application must carefully account for implementability, including transaction costs, signal timeliness, and robustness out of sample.

Ideas for future research follow naturally from these constraints and opportunities. First, richer NLP designs could move beyond one-dimensional “long/short” sentiment to separate tone, uncertainty, disagreement, and narrative themes (e.g., using topic models, embedding-based clustering, or event extraction). Second, future work could integrate YouTube sentiment with other complementary signals (news, macro variables, options-implied measures, flows, and social-media attention) in multivariate or machine-learning frameworks to evaluate whether the YouTube component adds incremental predictive power. Third, robustness and causality could be addressed more directly by exploiting quasi-experimental timing variation (upload shocks, breaking-news

events, or platform-level changes) and by testing lead-lag structure at finer frequencies (daily or intraday). Finally, from an applied perspective, future work could explore portfolio-construction and risk-management use cases, such as volatility timing, drawdown control, or regime detection, where sentiment signals may deliver more stable value than pure return forecasting.

Overall, the thesis suggests that YouTube is not just a consumption platform but also a measurable sentiment channel: when processed at scale, it can provide a novel lens on investor narratives and expectations, and it offers a promising direction for both academic research and systematic investment applications.

Bibliography

Ahmad, K., Han, J., Hutson, E., Kearney, C., & Liu, S. (2016). Media-expressed negative tone and firm-level stock returns. Journal of Corporate Finance, 37, 152–172. DOI: 10.1016/j.jcorpfin.2015.12.008

Anese, G., Corazza, M., Costola, M., & Pelizzon, L. (2023). Impact of public news sentiment on stock market index return and volatility. Computational Management Science, 20(1), Article 20. DOI: 10.1007/s10287-022-00462-6

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. 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. arXiv preprint arXiv:1908.10063. (Available at arXiv:1908.10063)

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

Ballinari, D., & Maly, J. (2025). FX sentiment analysis with large language models. (Swiss National Bank Working Paper No. 2025-11). Available at SSRN: http://dx.doi.org/10.2139/ssrn.5406332

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

Carhart, M. M. (1997). On persistence in mutual fund performance. Journal of Finance, 52(1), 57–82. DOI: 10.1111/j.1540-6261.1997.tb03808.x

Chen, P., Boukouvalas, Z., & Corizzo, R. (2024). A deep fusion model for stock market prediction with news headlines and time series data. Neural Computing and Applications, 36, 21229–21271. DOI: 10.1007/s00521-023-08665-3

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2025). Market signals from social media. SSRN Working Paper. Available at SSRN: https://ssrn.com/abstract=5187350 (Working paper, March 2025)

Divernois, M.-A., & Filipović, D. (2024). StockTwits classified sentiment and stock returns. Digital Finance, 6, 249–281. DOI: 10.1007/s42521-023-00064-w

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56. DOI: 10.1016/0304-405X(93)90023-5

Gabrovšek, P., Aleksovski, D., Mozetič, I., & Grčar, M. (2017). Twitter sentiment around the earnings announcement events. PLOS ONE, 12(2), e0173151. DOI: 10.1371/journal.pone.0173151

Garcia, D. (2013). Sentiment during recessions. Journal of Finance, 68(3), 1267–1300. DOI: 10.1111/jofi.12027

Heston, S. L., & Sinha, N. R. (2016). News versus sentiment: Predicting stock returns from news stories. (Finance and Economics Discussion Series Working Paper 2016-048). Washington, DC: Board of Governors of the Federal Reserve System. Available at: https://doi.org/10.17016/FEDS.2016.048

Kamal, S., Sharma, S., Kumar, V., Alshazly, H., Hussein, H. S., & Martinetz, T. (2022). Trading stocks based on financial news using attention mechanism. Mathematics, 10(12), 2001. DOI: 10.3390/math10122001

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, 33, 171–185. DOI: 10.1016/j.irfa.2014.02.008

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. Finance Research Letters, 62, 105227. DOI: 10.1016/j.frl.2023.105227

Kirtac, K., & Germano, G. (2024). Enhanced financial sentiment analysis and trading strategy development using large language models. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA), 1–10. (Association for Computational Linguistics)

Kirtac, K., & Germano, G. (2025). Sentiment-augmented reinforcement learning for portfolio optimisation with large language models (SAPPO). In Leveraging LLM-Based Sentiment Analysis for Portfolio Optimization (REALM 2025 Workshop). Available on OpenReview: https://openreview.net/forum?id=JBSTlpUUpm

Lefort, B., Benhamou, E., Ohana, J.-J., Saltiel, D., Guez, B., & Jacquot, T. (2024). Stress index strategy enhanced with financial news sentiment analysis for the equity markets. arXiv preprint arXiv:2404.00012. Available at arXiv: https://arxiv.org/abs/2404.00012

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

Lopez Avila, E., Martineau, C., & Mondria, J. (2024). Social media and the distortion of price revelation. SSRN Working Paper (May 16, 2024). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4439793

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015) (pp. 1357–1362). IEEE. DOI: 10.1145/2808797.2809350

Shen, S., Xia, L., Shuai, Y., & Gao, D. (2022). Measuring news media sentiment using big data for Chinese stock markets. Pacific-Basin Finance Journal, 74, Article 101810. DOI: 10.1016/j.pacfin.2022.101810

Si, J., Mukherjee, A., Liu, B., Pan, S. J., Li, Q., & Li, H. (2014). Exploiting social relations and sentiment for stock prediction. In Proceedings of EMNLP 2014 (pp. 1139–1145). Association for Computational Linguistics.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. European Financial Management, 20(5), 926–957. DOI: 10.1111/j.1468-036X.2013.12007.x

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. 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. Journal of Finance, 63(3), 1437–1467. DOI: 10.1111/j.1540-6261.2008.01362.x

Warkulat, S., & Pelster, M. (2024). Social media attention and retail investor behavior: Evidence from r/wallstreetbets. International Review of Financial Analysis, 96, 103721. DOI: 10.1016/j.irfa.2024.103721

Watanabe, M. (2025). Social media sentiment and over-the-counter equity returns. Unpublished working paper, June 30, 2025. (Manuscript available via Keio University seminar series)

Zhang, P., Harris, R. D. F., & Zheng, J. (2024). GNN-based social media sentiment analysis for stock market forecasting and trading. SSRN Working Paper No. 5051165. Available at SSRN: https://ssrn.com/abstract=5051165

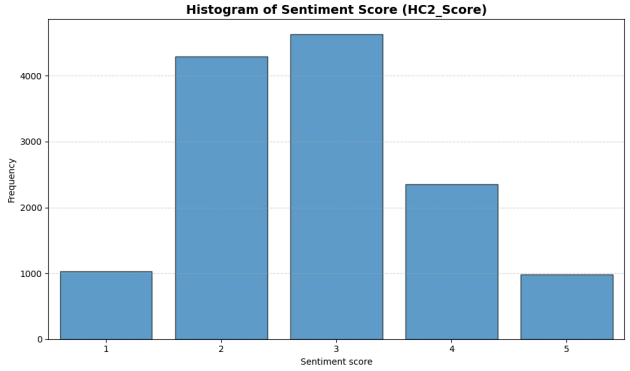
Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM) (pp. 375–378). AAAI Press.

Appendix 1: Sentiment Data



Summary statistics of sentiment score (EN2_Score)

	Value
Count	48628
Mean	3.07
Std dev	0.93
Skew	0.27
Kurtosis	-0.38



Summary statistics of sentiment score (HC2_Score)

	Value
Count	13277
Mean	2.85
Std dev	1.04
Skew	0.29
Kurtosis	-0.46



Summary statistics of sentiment score (TN2_Score)

	Value
Count	102577
Mean	3.38
Std dev	0.96
Skew	0.01
Kurtosis	-0.76

Appendix 2: Regressions Results

For the Energy Industry (Ticker: XLE)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.025
Model:	OLS	Adj. R-squared:	0.009
Method:	Least Squares	F-statistic:	4.192
Date:	Tue, 06 Jan 2026	Prob (F-statistic):	0.0449
Time:	21:38:12	Log-Likelihood:	-166.58
No. Observations:	64	AIC:	337.2
Df Residuals:	62	BIC:	341.5
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	28.0914	13.581	2.068	0.043	0.942	55.240
x1	-9.0595	4.425	-2.047	0.045	-17.905	-0.214

Omnibus:	26.879	Durbin-Watson:	2.099
Prob(Omnibus):	0.000	Jarque-Bera (JB):	62.838
Skew:	-1.282	Prob(JB):	2.26e-14
Kurtosis:	7.122	Cond. No.	184.

For the Healthcare Industry (Ticker: XLV)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	10.16
Date:	Tue, 06 Jan 2026	Prob (F-statistic):	0.00384
Time:	21:47:07	Log-Likelihood:	-61.046
No. Observations:	27	AIC:	126.1
Df Residuals:	25	BIC:	128.7
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.6849	3.441	3.105	0.005	3.598	17.771
x1	-3.7125	1.165	-3.187	0.004	-6.112	-1.313

Omnibus:	1.515	Durbin-Watson:	2.471
Prob(Omnibus):	0.469	Jarque-Bera (JB):	1.210
Skew:	-0.501	Prob(JB):	0.546
Kurtosis:	2.729	Cond. No.	66.4

For the Technology Industry (Ticker: XLK)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.041
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	5.472
Date:	Tue, 06 Jan 2026	Prob (F-statistic):	0.0225
Time:	21:55:51	Log-Likelihood:	-179.41
No. Observations:	66	AIC:	362.8
Df Residuals:	64	BIC:	367.2
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	26.7802	11.128	2.407	0.019	4.550	49.011
x1	-7.2288	3.090	-2.339	0.022	-13.402	-1.055

Omnibus:	6.093	Durbin-Watson:	2.374
Prob(Omnibus):	0.048	Jarque-Bera (JB):	5.911
Skew:	-0.463	Prob(JB):	0.0520
Kurtosis:	4.137	Cond. No.	135.

Appendix 3: Use of AI

Artificial Intelligence played a limited, yet supportive role in this thesis. The tool used was ChatGPT 5.1. The large language model (LLM) was primarily used to enhance writing quality and clarity, enabling us to improve grammar, shorten lengthy sections and rephrase sentences to make them clearer.
The tool was not used to generate analysis or theoretical insights, but rather refine these once concluded to assure a better flow of text, and any output has been reviewed thoroughly

De-Biasing Venture Investments Through Alternative Data

Index

Introduction 3

1. Venture Capital Decision Making 4

1.1. Overview of the Process 4

1.1.1. Pre-Investment 4

1.1.2. Deal Structuring 5

1.1.3. Post-Investment 5

1.2. Investment Selection 6

1.2.1. Success Factors 6

1.3. Biases in Investment Decisions 9

1.3.1. The Herding Bias 10

1.3.2. Mitigating Biases 12

2. The Use of Alternative Data in Investment Decisions 14

2.1. The Emergence of Alternative Data 14

2.1.1. Alternative Data to Forecast 14

2.1.2. Alternative Data in Venture Investments 15

2.2. Investor Sentiment 18

2.2.1. Investor Sentiment and Market Dynamics 18

2.2.2. Investor Sentiment and Venture Capital 22

2.2.3. Criticises in Modelling Market Sentiment 24

3. De-Biasing Venture Investment Decisions 26

3.1. The Current State of the Research 26

3.2. Hypothesis Formulation 28

3.3. Methodology 29
3.3.1. YouTube Sentiment Analyser 29
3.3.2. Derivation of Sentiment and of *Herding Index (HI)* 32
3.3.3. The Final Sentiment Dataset 36
3.4. A New Definition of Success in Venture Investments 39
3.4.1. Measuring Success 39
3.4.2. The Final Valuations Dataset 41
3.4.3. A comparison with Previous Metrics 44
3.5. Analysis of the Results 45
3.5.1. Preliminary Analysis 46
3.5.2. Regression Results 47
3.5.3. Interpretation of the Results 53
Conclusion 56
References 58
Appendix 1 67
Appendix 2 69
Appendix 3 70

Introduction

Investment selection is arguably the most important aspect of venture investing (Gompers, 2020), yet it is also the most challenging one, due to the highly uncertain environment in which investors operate. This environment has historically fostered the development of biases in decision making, particularly evident when investors face new technological advancements, whereby markets typically display signs of *irrational exuberance* (Greenspan, 1996), a phenomenon observed during every major technological shift.

In the course of this thesis, we dive deep into the venture capital space through the lens of behavioural finance. In the first chapter, we explore these two worlds and their relationship by reviewing the current state of the literature. In the second chapter, we gain a deeper understanding of how these two fields are financially modelled. Finally, in the third and concluding chapter, we develop a new model to study their relationship, introducing substantial innovations both in the ways in which these two fields are commonly thought by themselves, and in the way in which they interact. To do so, we focus on the most prominent behavioural bias in the venture capital space: the herding bias.

We develop a new model to quantify the amount of herding in the venture capital market and find that there exists a significant negative relationship between the quality of venture investments and the amount of herding registered in the market, especially in times of extremely high or low returns. This suggests that having a contrarian opinion, and holding it even in turbulent times, could really pay off in the long term.

In this context, rather than aiming at making human judgment obsolete—by automatizing investment selection as many papers and funds attempt to do—the use of data and artificial intelligence present an opportunity to develop tools that enhance (and not replace) the decision-making processes in a model of hybrid intelligence (Dellermann et al., 2021), which literature suggest could be particularly useful for less experienced investors (Blohm et al. 2020).

1. Venture Capital Decision Making

The venture capital decision-making process is a complex procedure that involves identifying, evaluating, and winning the trust of high-potential startups. Being successful in this business does not only mean having grasped the typical startup evaluation rules, but also developing a deep understanding of products, technologies, industries, market trends, psychology of the founders and much more. The ability to comprehend and correctly prioritize all these factors is crucial for identifying startups that can deliver exceptional returns in the long term and create a successful exit (Kaplan, Sensoy, Stromberg, 2009).

The importance of having these transversal skills and knowledge, frequently means that venture capital investors, compared those of other, more established asset classes, operate in a highly uncertain environment, requiring them to balance intuition, analysis, and experience when making investment decisions (Gompers & Lerner, 2001).

1.1. Overview of the Process

Venture capital firms do not necessarily follow strict guidelines for structuring their decisions. However, on a general level, we can state that this is typically divided into three main phases: pre-investment, deal structuring, and post-investment (Gompers et al., 2020). Each phase is essential to the overall success of the investment and involves a series of activities needed for selecting, supporting, and exiting their investments successfully.

1.1.1. Pre-Investment

The pre-investment phase is the first, essential step that serves as the foundation of the venture capital process, where the firm focuses on sourcing and selecting the most promising businesses or entrepreneurs to invest in. In this phase, the typical activities include market research, identification of emerging technologies, evaluation of potential investment opportunities (pitches), conducting thorough (or not) due diligence, and assessing the management team's capabilities to build and run successfully the startup. The investor needs to analyse the business model, assess the product-market fit, investigate the competitive landscape, and forecast its size and future evolution. This

phase is particularly important, as it sets the stage for the entire investment lifecycle. It is vital, for a fund, to find and invest in the best startups so that they can return enough capital to LPs and create an upside for themselves (GPs)[1].

1.1.2. Deal Structuring

Once a potential investment is identified, the firm enters the deal structuring phase, where the focus shifts to materially structuring the deal and negotiate over the terms of the investment, if necessary. In this phase, the investment committee of the fund typically takes a decision on whether to invest or not in the startup, and then presents to it a term sheet[2], upon which the two parties will need to agree on the legal and financial aspects of the deal, such as the valuation, the ownership percentage, the board representation and other governance rights. Here the investor needs to balance very carefully its interests with those of the startup to ensure a mutually beneficial agreement, and to avoid losing the deal over some other fund that might have also presented (or will present soon) another term sheet. Effective negotiation during this phase is essential to protect the firm's investment and lays the groundwork for future collaboration. After all, the startup and the firm will need to partner for many years ahead and it is extremely important that they get along well with each other.

1.1.3. Post-Investment

When the deal is closed, the firm moves into the post-investment phase, where the focus becomes actively managing the investment to maximize its value. Venture capitalists make few investments a year and tend not to build fully diversified portfolios. This means that actively reducing its risk profile is strategic and widespread practice for a venture capital fund. In this, funds tend to differ a lot, as some tend to help more (sometimes even by being too intrusive) while others might give complete freedom to the founders to operate.

To help the portfolio company, the fund closely monitors the startup's performance, provides strategic guidance (frequently through a series of advisors), leverages the firm's network to aid the business development function or attracts talent that can join the

[1] Limited Partners (LPs) are the individuals or entities that invest in the fund, General Partners (GPs) are the ones that manage it. They both participate in the returns of the fund.
[2] A document summarizing the contractual terms of the investment.

startup, thus helping the portfolio company in navigating its early stages of growth. The venture capital fund may also take a more active role in shaping the startup's strategic direction, setting up its operations and preparing it for future funding rounds or an eventual exit.

Ultimately, the goal of the post-investment phase is to achieve a successful exit, whether through an IPO, M&A, or another liquidity event, where the firm realizes the return on its investment.

1.2. Investment Selection

Through the above-mentioned interconnected phases, venture capital firms work to maximize the value of their investments, with the pre-investment phase being particularly important as it lays the groundwork for long-term success. Inside this phase, the most critical (and most difficult) task is to select the most promising investment (Gompers et al., 2020).

1.2.1. Success Factors

To make these extremely complex decisions, investors tend to follow a great variety of diverse processes. However, generalizing, we can state that the evaluation process of an early-stage startup revolves around three core aspects of the startup: the founding team, the market potential, and the product or idea envisaged[3]. Each of these features provides essential insights into the startup's upside potential and helps VCs identify the most promising early-stage investments.

Founders

The first element every VC assesses is the founding team, often considered the most critical element of a startup's potential (Sudek, 2006). VCs look beyond just the resumes of the founders; they really delve deep into the team's track record, leadership qualities, and personal drive. Some VCs look for a specific mix of features in the founding team, whereby this needs to contain both the problem that the startup is trying to solve, and therefore someone coming from this domain, and the solution, meaning someone with an idea and the skills to tackle that specific problem (Forbes, 2020).

[3] In later stages, financials will become more reliable and thus more used in assessments.

Experience is another key factor—founders who have successfully navigated the startup journey before and had experience in fast growing, unstructured environments or that possess deep expertise in their industry are viewed more favourably[4]. However, experience alone is not enough; VCs also want to see visionary leadership, the ability to inspire and rally a team, and a relentless drive to overcome obstacles (Franke, Gruber, Harhoff, & Henkel, 2008). In a way, they want to see someone with the potential to revolutionize entire industries. They judge the founders' problem-solving skills, resilience, and adaptability—traits that are of extreme importance in the unpredictable world of startups. A strong founding team can more easily manage pivots[5] and innovate in response to market feedback, which is often crucial for the success of the venture (Bock & George, 2018).

Additionally, VCs consider the team's cohesion and chemistry (also known as *the founder's mix*). A founding team that works very well together, communicates efficiently, and shares the same vision for the future of the company is obviously more likely to survive the difficulties of building a business (Klotz et al., 2014).

Finally, investors, especially if they prefer to take an active role in the development of the startup, also want to find founders who are coachable—those who are open to feedback and willing to learn from their advisors and mentors or, perhaps even more importantly, from the market itself.

Ultimately, VCs invest in people as much (arguably more) as in ideas, and they want to back founders who have the passion, skills, and determination to turn a vision into reality.

Market

The market is the second crucial aspect that VCs evaluate. Even the best product or the most talented founding team will fail if the market opportunity is not big enough (Aerts et al., 2007). VCs try to find startups that are targeting large and growing markets, where there is great demand for new, innovative solutions. Market size is the most essential element to keep into consideration; VC-backed startups must have the possibility to capture a significant share of a what needs to be a multi-billion-dollar market. This great

[4] Something that is particularly appreciated are *Serial* Founders, meaning founders who founded (and possibly exited) several startups in their lives.
[5] Change in the company's strategy and business model, product, or target market.

potential allows the company to have enough space to grow and scale, which is of extreme importance for creating those outsized returns that VCs are so hardly looking for (Yin and Luo, 2018).

Other than the size of the market, it is essential for VCs to evaluate the growth trajectory of the market. After all, if the market is growing, it means that it is easier to seize it. In fact, markets that are expanding rapidly offer more opportunities for new entrants to easily establish themselves and gain significant market share (Gruber, MacMillan, & Thompson, 2008).

Additionally, VCs tend to analyse the competitive landscape as a whole, in order to understand the level of competition and the uniqueness of the need that the startup is trying to tackle. After all, ideally, every startup needs to become a monopoly in its own niche (Thiel, 2014), especially if the market is new or does not even exist yet. Moreover, VCs are particularly interested in markets that are fragmented or undergoing significant disruption, as these present great opportunities for startups to innovate and capture market share from more established players After all, as long as startups will get to innovation before incumbents get to distribution, they will be able to grow successfully (a16z, 2015).

Finally, one last key aspect of market evaluation is the identification of the startup's target customer segment. Here, investors want to see a deep understanding of who their ideal customer is, what are its needs and pain points, as well the marketing and commercial strategy that the startup is going to use to attract, win and retain those clients[6]. A clear path to market adoption, either through a unique go-to-market strategy or a compelling value proposition, is crucial. Start-ups that can demonstrate an existing market presence - such as initial customer acquisition, revenue growth or strong user engagement - are particularly attractive to VCs, especially if they can show this in their financials.

Product

The third and final element that investors need to analyse is the product itself. The product is the tangible manifestation of the venture's vision and the primary means by which it will capture part of the market value. VCs assess the uniqueness of the product

[6] In this context, creating an Ideal Customer Profile (ICP) can be extremely useful.

and its differentiation from existing solutions (Yin and Luo, 2018). The literature suggests that this may even be the most significant element in determining the success of a startup, as noted by Kaplan et al. (2009), but it can be difficult to recognise, especially in the very early stages of the venture.

Venture capitalists are interested in opportunities that offer an interesting value proposition that can solve real problems for customers in a significantly better way than existing alternatives (if any). Whether it is a breakthrough technology, a new business model or a unique set of features, the product must have a clear competitive advantage (Gans, Hsu and Stern, 2002).

In assessing the product's potential, VCs place significant attention to its scalability. This is another critical factor in product evaluation. Venture capitalists want to see that the product can scale efficiently as the company grows. This involves assessing the underlying technology, the architecture of the product, and the venture's ability to scale and deliver the product without compromising quality or performance. By definition, a scalable product is one that can serve an increasing number of customers or expand into new markets with relatively low additional costs (Schwienbacher, 2013).

Finally, product-market fit is key. Venture investors need to find significant evidence that the product resonates with its target audience, which can be demonstrated through customer feedback, pilot projects or early commercial traction. A product that has achieved strong market fit typically shows signs of organic growth, often signalled by the spreading of word-of-mouth referrals or high customer retention rates. Ventures that have developed a clear product roadmap[7] and a strategy of continuous improvement and innovation are particularly attractive as they tend to signal the potential for long-term success (Thiel, 2014).

1.3. Biases in Investment Decisions

When making investment decisions, investors often operate in a context of extreme uncertainty and urgency. These two factors significantly increase the likelihood of falling prey to cognitive biases (Bekiros et al., 2017).

[7] A plan for the development of the product's future features.

As evidenced by Barber and Odean (2001), uncertainty may bring overconfidence, which can be defined as an individual's tendency to overestimate their own abilities, knowledge, or the accuracy of their predictions. This bias works in parallel with the herding bias in the causations of the ups and downs characterizing this market. This, very recently, has also been found to be related with a longer time to exit (Ben Amor & Kooli, 2024).

Huang & Pearce (2015) further find that investors tend to rely on *gut feelings* instead of properly structured analysis. Surprisingly, this is found to be a successful way to determine good investments, but only for seasoned investors and better if accompanied by some amount of more structured analysis.

However, the most important bias of all, for its severe consequences on the industry[8], is herding.

1.3.1. The Herding Bias

A well-documented occurrence in various financial markets is the *herding behaviour*, a phenomenon whereby investors mimic the actions of others rather than relying on their own personal beliefs, disregarding the information they possess (Banerjee 1992).

In the venture capital industry, as just mentioned, herding behaviour can be significant and leads the industry to be characterised by strong booms and contractions.

Evidence of Herding in Venture Capital

One of the most common expressions of herding behaviours is the so-called "valuation bubble." Here, the excessive investment in certain sectors, companies, and enterprises usually leads to valuations that result inflated if looked at in the short term. This was particularly evident during the dot-com bubble of the late 1990s and early 2000s, where a rush to invest in internet startups led to unsustainable valuations (Gompers & Lerner, 2001); a true example of herding behaviour among VCs that contributed to this bubble, as the ever-present fear of missing out (FOMO) on the next big occasion led to overinvestment in highly conjectural endeavours.

Furthermore, it is likelier that self-assured VCs will choose to invest in startups with poor prospects or will tend to overlook critical risks, which will then, as a result, lead to higher

[8] FOMO (Fear of missing out) and herding are regarded as the two greatest dangers in investor's mentality, frequently driving investments that should not have been made (CBInsights, 2023; The Guardian, 2023).

rates of failure in their portfolios (Zacharakis & Shepherd, 2001). This problem not only exerts influence on the returns of the VCs themselves but can also have larger implications for the startup ecosystem; and this happens because failed ventures absorb funds that could have been allocated to opportunities that were more viable. This is commonly referred to as resource misallocation, where capital is condensed in a few highly visible startups, while other potentially feasible ventures struggle to secure funding; this can suppress innovation as the diversity of investment opportunities is diminished (Nanda & Rhodes-Kropf, 2013).

Another structure that can be disrupted by herding is the ability to make a good exit on the investment. When many VCs simultaneously rush to exit an investment, it can lead to sharp declines in valuations and this, in turn, can have the effect of negatively impacting not just the targeted company but also the market in a broader way. Zhang et al. (2014) find that herding, even though it can increase the short-term valuations of new ventures, is found to have an adverse effect on the long-term valuations which in turn hinders the performance of investors. This is coherent with the idea that agents, when under uncertainty and asymmetric information, do not select the most plausible assets based on their fundamentals; however, they do tend to opt for assets with high market values in the short term (Akerlof & Shiller, 2009; Siemer, 2010; Sewell, 2011).

Causes of Herding in Venture Investments

As previously mentioned, uncertainly and urgency can be thought of as the main drivers of herding in the markets. In fact, in public markets information about firms is relatively accessible, simple and clear, whereas in private ones, especially in the venture capital space, even the simplest piece of information is difficult to acquire, and strong information asymmetries between founders and investors make this information not only hard to find, but frequently unreliable. This difficulty to obtain information can lead VCs to approach a heuristic method for decision-making following the actions of more informed peers.

Nonetheless, there exist other causes of this phenomenon.

Reputation concerns play a significant role in driving herding behaviour. According to Dasgupta, Prat, and Verardo (2011), particularly less experienced VCs, may take part in herding to protect their reputations. By associating their investment decisions with those

of more established VCs, they can show competence to other market participants. “Social learning” is a very crucial concept where venture investors perform the actions of others as implicit endorsements of the investment opportunity, leading to a bandwagon effect (Banerjee, 1992).

In the VC community, another cause of the investment decisions is the network effects. In fact, VCs often operate in tight-knit networks, where relationships and reputation are crucial. Studies like those by Hsu (2004), Hochberg, Ljungqvist, and Lu (2007) or Zava and Caselli (2023) show and explain very clearly that VCs with strong network ties are more likely to invest in the same deals as their network peers. These network effects can aggravate herding behaviours, as VCs may give maintaining relationships priority over independent analysis.

Finally, some other studies like Scharfstein and Stein (1990) emphasise that this behaviour could actually be rational from an individual investor’s point of view, as emulating the investment choices of more experienced or better-informed investors can reduce the risk of adverse selection.

1.3.2. Mitigating Biases

Considering the significant impact that herding and other biases have, various strategies have been proposed to mitigate its effects.

Studies by Forbes and Milliken (1999) indicate that teams with diverse backgrounds and perspectives are less likely to fall prey of overconfidence, as differing viewpoints can challenge overly optimistic assumptions and promote a more balanced decision-making[9]. Moreover, structured decision-making processes and the use of objective criteria in investment evaluations can help reduce the influence of overconfidence. Research by Kahneman and Lovallo (1993) suggests that, yet again, incorporating external perspectives and conducting pre-mortem analysis can also be effective in countering overconfidence. Kahneman (2011) further develops the idea that, by simply acknowledging the risk, the human mind can already contrast biases for the great part[10].

[9] This is the reason diversity (at GPs level) is truly one of the most important assets a venture capital firm can have.
[10] This idea will be picked up again later and can be considered one of the pillars of this work.

In more recent times, Blohm et al. (2020) discovered that if on the one hand venture investors with significant investment experience can mitigate biases such as overconfidence, loss aversion and local prejudice leveraging their knowledge, less experienced ones can be affected by these biases, which significantly influence investment performance.

Ultimately, grounding opinions in objective data, could really help avoiding common pitfalls in decision-making. This is, as we will see in the next chapter, the direction that the industry has taken in the last few years.

2. The Use of Alternative Data in Investment Decisions

Alternative data can be defined as non-traditional data sources offering unique insights that go beyond the conventional data typically found in decision-making processes. Unlike financial statements, economic indicators, and other sources of standard data, alternative data encompasses a wider array of information including social media activity, satellite imagery, transaction data, and traffic on the internet.

Businesses and investors can gain real-time, granular, and predictive insights through these unconventional sources of data; and this is useful to uncover different trends and patterns, obtaining a competitive edge. In this sense, the use of alternative data is more than ever reshaping how decisions are made across industries, and only 5% of it has been turned into useful information, as the main difficulty is to extract a valid and readable structure from it (Eagle Alpha, 2024).

2.1. The Emergence of Alternative Data

In the past decade, the use of alternative data in investment decisions has significantly grown, driven by advances in technology and data analytics. Traditional financial data, like earnings reports and balance sheets, are often too old[11] and could be easily manipulated. On the contrary, alternative data can offer real-time insights and can provide better and deeper comprehension of company performances and market trends (Bollen, Mao, & Zeng, 2011).

2.1.1. Alternative Data to Forecast

Throughout the investment space, alternative data is primarily used for enhanced forecasting and risk management. Hedge funds and asset managers are increasingly leveraging data from sources such as credit card transactions, satellite imagery, and social media; this is done to determine consumer behaviour, monitor supply chains, and track economic activities (Kolanovic & Krishnamachari, 2017). A simple example is that of

[11] As most of the information contained in the figures often comes too late to take any relevant decision.

using satellite imagery to estimate retail foot traffic or to monitor the construction progress of large infrastructure projects; by doing so, valuable insights that are not immediately reflected in financial statements can be provided.

One of Tetlock's early studies (2007) highlighted how media sentiment could forecast significant changes of stock prices, creating a situation in which a larger adoption of alternative data in the equity investments field can happen. Furthermore, many investors have been encouraged to incorporate these insights into their models by the availability of vast amounts of data from non-traditional sources with the hope of gaining an edge over competitors who may still use exclusively conventional sources of data.

It is in this context that sentiment analysis from social media and news sources has become a novel and valuable tool for equity investors. Through the analysis of large volumes of text data, investors can assess market sentiment and possibly predict short-term stock price movements (Renault, 2017). This is particularly useful for multiple cases such as anticipating market reactions to corporate announcements, political events, or macroeconomic changes.

Another critical application is the identification of early signals of company performance. For example, analysing web traffic data can reveal trends in consumer interest before they are reflected in sales figures or earnings reports. Through this proactive approach, investors can position themselves ahead of market shifts, discovering the potential of higher returns (Gupta, Mudgill, & Wang, 2019).

2.1.2. Alternative Data in Venture Investments

In the venture capital space, where the availability of reliable financial data is often extremely limited, alternative data is starting to find its first applications in the above-mentioned pre-investment, deal structuring, and post-investment phases. By providing insights that are not available through traditional data sources, alternative data can help VCs make better and more informed investment decisions, reducing risk and identifying high-potential startups earlier in their lifecycle (Bonein & Carat, 2020).

When sourcing new investment opportunities, VCs can analyse web traffic, app downloads, and social media engagement with the goal to identify startups that are gaining traction with customers, even before they come in the spotlight of the broader

investment community (Gompers, Gornall, Kaplan, & Strebulaev, 2020). This is key in allowing venture investors to discover promising startups at earlier stages, giving them a competitive advantage in securing these sought after deals.

In the monitoring phase, alternative data provides VCs with a better and more comprehensive view of a startup's operations and its market position. For instance, analysing employee reviews on platforms like Glassdoor or monitoring customer feedback on other social media can offer valuable insights into a startup's own organizational health and showcase product reception that may not be so evident if we only look at the traditional financial metrics (Sahut, Boulerne, & Teulon, 2019).

Challenges

Despite the growing importance of alternative data, its use in equity and venture investments is subject to multiple challenges. Although the venture capital industry is remarkably close to the world of technology and innovation, the integration of alternative data sources into already existing investment models can be difficult.

Indeed, traditional investment models are often not designed with the aim of incorporating the vast and unstructured nature of alternative data. Developing the necessary infrastructure and expertise to analyse, and subsequently integrate this data, can be resource-intensive (Kolanovic & Krishnamachari, 2017). Moreover, the sheer volume of alternative data may very well lead to information overload, losing the key signals in a vast sea of noise.

The rise of Data-Driven VC

Despite these criticisms, the advent of AI has recently given new life to the so-called *data-driven venture capital*[12] phenomenon, creating a significant impact in the entire industry and paving the way for innovation.

Theoretically, if one could access sufficient and meaningful data to evaluate a startup across the three critical dimensions—founders, market, and product—it should be possible to conceive a formula able to predict, in advance, the potential return of each

[12] A whole community has been built around this topic, see https://www.datadrivenvc.io/.

startup investment. This concept is precisely what the emerging class of *data-driven* investors aims to achieve (Corea et al., 2021) [13].

Numerous studies have identified the most key factors contributing to a startup's success, often by training machine learning (ML) models, and in particular artificial neural networks (ANNs), on vast datasets. Research has recognized several key factors that are strictly bound to the dimensions already explored in the previous chapter[14]. Some of the most important ones are the total amount of funding, industry focus, product description, employee count, year of founding, founding team size, founders' previous experience, website traffic, social media presence, number of patents, and finally, total capital raised (e.g., Caruso et al., 2015; Cao et al., 2022; Ross et al., 2021; Corea et al., 2021).

Other studies suggest that qualitative data can also be used with the aim of training machine learning models, giving significant results (Bai & Zhao, 2021), and pointing to the possibility for even more unconventional data to be successfully integrated into venture investment decisions.

Hybrid Intelligence

Some scholars advocate for an approach that integrates human judgment with machine intelligence, rather than replacing one with the other; an approach that, by doing so, is perceived as more balanced. This is the *hybrid intelligence* model, proposed by Dellermann et al. (2021), and it emphasizes the critical role humans play in handling *soft* information and navigating situations that are often characterized by unknowable (and non-modellable) risks.

Machines encounter limitations in two fundamental areas: in the processing and interpretation of soft information—data that cannot easily be quantified—and in making accurate predictions for scenarios of extreme uncertainty; scenarios where the available information does not adequately represent the multiple potential outcomes. It is in these

[13] This of course assumes the existence of a common denominator among all the successful venture investments. Arguably, if we believe that one of the few things that all highly successful startups have in common is their ability to discover value in unexpected places by approaching business from first principles, rather than relying only on already accepted formulas (Thiel, 2014), then this whole movement would lack its fundamental assumption and would eventually either become obsolete or find a way to model the *innovativeness* of an idea.

[14] See 1.2.1 Success Factors.

contexts that the addition of the human intuition and judgement remains essential for the assessment of signals that machines are not able to fully able to grasp.

This model will be of inspiration for the rest of the thesis, as it is a third way to follow that avoids both the biases to which purely gut feeling-oriented investors are subject to but also avoids the impossibility to find, and invest, in ideas that are completely out of the current schemes, which a machine learning model would not be able (yet) to identify.

2.2. Investor Sentiment

Behavioural finance has historically found its most prominent source of interest in investor (or market)[15] sentiment, which has always been considered a crucial factor in the stock market, due to its strong ability to influence investor behaviours, prices, and overall market dynamics. It mirrors the collective mood or attitude of investors towards market conditions, individual stocks, or broader economic factors. Ultimately, the process of modelling market sentiment implies analysing various data sources to quantify these attitudes and leverage them to predict market movements.

In this field, research has been focusing in the last few years on two distinct aspects: understanding how the market is influenced by sentiment and finding ways to derive sentiment from alternative data sources[16].

2.2.1. Investor Sentiment and Market Dynamics

Research has explored how sentiment influences markets, focusing on the influence it has on returns and in the causation of extreme market conditions.

Sentiment and Returns

In fact, investor psychology, influenced by external factors such as news events, economic data releases, and broader market trends, can lead to overreactions, underreactions, or herd behaviours, all of which are heavily impacting stock prices (Baker & Wurgler, 2007). Consequently, it becomes crucial for investors to understand and model sentiment, as it can provide great insights into potential market mispricing and investment opportunities.

[15] In the resto of the thesis these two terms will be used interchangeably.
[16] As we will see, this thesis contributes substantially to both these fields.

Tetlock (2007) demonstrates this relationship by analysing media (journals) sentiment and showing that it has a significant impact on stock market returns, clearly showing that the tone of Wall Street Journal columns can predict stock market movements. Bollen, Mao, and Zeng (2011) confirm this theory by analysing not media, but *social* media sentiment. They found, yet again, a significant relationship with stock returns and the possibility to predict them before they realized.

Recently, Deveikyte et al. (2022) further investigate Twitter sentiment and find that it is, to their surprise, negatively related to market volatility.

If these results hold, then, to summarize them in a sentence, we could say that if people are happy price levels tend to increase (positive relationship between sentiment and returns), and the volatility of the returns tends to decrease (negative relationship between sentiment and returns' volatility). This is coherent with the so-called *leverage effect* (Black, 1976), whereby volatility rises when markets collapse but not as much when markets boom.

Finally, Li et al. (2023) found that there is a significant correlation between social media sentiment and investor sentiment (more precisely, retail investors). This means that all previous studies, when measuring media or social media sentiment, were correctly capturing what investor were thinking.

Sentiment and Extreme Market Conditions

Shiller (2000) was the first to recognize the importance of examining sentiment to more comprehensively understand the broader market dynamics. He argues that sentiment can be a disruptive phenomenon that provokes asset bubbles and crashes, given that oftentimes prices are driven more by investors' emotions than by fundamental value.

Wang & Wang (2018) build on top of this and develop a model in which *gurus* influence the market's sentiment to study the effects on volatility in market returns. They discover that during periods of market booms or busts, investors are more likely to herd (follow) around these influential individuals. As a consequence of this, especially in poorly informed markets, such as the venture capital one, this tendency can lead to decreased volatility, as collective investor behaviour becomes more synchronised, other than obviously extreme movements in the value of assets.

Further, Ren and Wu (2018) link these results back to investor sentiment. In their study they introduce a new framework that can be used for identifying herding behaviours by analysing the convergence or divergence of investor opinions, proxied by sentiment observations. This approach suggests that by examining how similar or varied investors' sentiments are, it is possible to detect herding in the market more efficiently and potentially even more accurately than any traditional method based solely on asset prices. Linking this result back to the venture capital space, where timely and accurate data is scarce, sentiment analysis could prove extremely useful in revealing herding behaviours in the market.

Sentiment and Herding Bias

Herding behaviours, which are deeply rooted in human psychology and part of social activities, have traditionally been detected mainly through the analysis of asset prices, as discussed in the literature (Chang, Cheng, & Khorana, 2000; T. C. Huang, Wu, & Lin, 2016).

Arguably, this traditional method might present some disadvantages. For example, it suffers from strong delays. This is because market data is a consequence of investor decision-making behaviours, which may lag behind the actual onset of herding if this takes time to manifest explicitly. In that case, not having any way to verify if there is herding in the market, could lead to potential disasters as history demonstrates. Moreover, this delay is even more pronounced in markets that are not efficient[17] and have a limited amount of observable data, and this is the case for venture capital markets.

Given these strong limitations, detecting herding behaviours through sentiment analysis may provide a more timely and relevant approach. Recently, this approach has been favoured.

Modelling Market Sentiment

Modelling market sentiment is extremely difficult and relies on a variety of alternative data sources and analytical techniques. The most used methods are textual analysis, machine learning, and quantitative models.

The first of these methods, textual analysis, involves leveraging large volumes of text data from various sources (news articles, industry or financial reports, social media, and many

[17] Markets do not reflect all the available information.

others) to extract information about the writer's sentiment. This approach frequently uses natural language processing (NLP) techniques that can distinguish between positive, negative, or neutral sentiment contained in the text.

In recent years, machine learning models have increased in popularity when it comes to analysing sentiment. These models can be trained on labelled datasets to recognize patterns in the text and predict sentiment with a surprisingly high accuracy. Techniques such as support vector machines, neural networks, and deep learning, have been used extensively to examine not only social media data or news sentiment but even earnings call transcripts, which were demonstrated predictive of future stock prices (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014).

Quantitative models allow for a mixture of sentiment analysis and traditional financial metrics to create predictive models of stock market returns. The purpose of these models is to integrate sentiment scores from textual analysis with financial ratios, trading volumes, and other market indicators to forecast price movements or volatility in an even more accurate manner (Schumaker & Chen, 2009).

The data needed to develop these modes is frequently taken from alternative data sources such as social medias. This is why sentiment analysis has gained significant traction with the rise of platforms like Twitter, Reddit, and StockTwits. Investors are increasingly more reliant on the real-time nature of social media to gauge market sentiment, particularly when it comes to retail-driven stocks or during market-moving events (Peakmetrics, 2024).

Finally, it is important to note how some studies not only built models, but also created tools that can be of use to them. One example of these tools is the financial sentiment dictionary that was developed by Loughran and McDonald (2011), who grouped the most widely used expressions and words in the field. In this way, they allowed textual analysis tools to categorise words according to their sentiment in a financial context, capturing the real meaning of all the words, expressions and *slang* that rotates around the world of financial markets.

2.2.2. Investor Sentiment and Venture Capital

Investor sentiment is particularly relevant in the venture capital space due to the elevated level of uncertainty and the lack of traditional financial data available for startups. The literature has little developed this field, and some areas of it have remain untapped to this date[18].

Venture capitalists tend to rely (arguably too much) on market sentiment as a proxy for the potential success of an investment. Positive sentiment can drive up valuations and attract additional investors, while negative sentiment can lead to decreased interest and a harder fundraising journey (Gompers et al., 2008).

Moreover, research by Brown et al. (2012) have demonstrated how market sentiment can impact venture capital flows, whereby periods of high sentiment tend to be associated with increased investment activity.

Finally, Tumasjan et al. (2021) studied how Twitter sentiment can influence the judgement of venture investors and found that, if incorporated in their judgement, it can lead to a certain level of over-optimism about certain technologies, which hints to the existence of an overconfidence bias. This heightened optimism, emerging from the sentiment polarity[19], is discovered to be correlated with lower long-term returns; essentially, the more positive the sentiment around a certain technological advancement, the poorer the eventual investment outcomes may be.

Modelling Sentiment in the Three Phases of the Venture Capital Process

Modelling investor sentiment in the venture capital industry involves the collection and analysis of various data sources, including both traditional and alternative ones. Several methodologies have been created to quantify and analyse sentiment in this specific space, especially textual analysis tools and machine learning models. Additionally, more traditional approaches remain very used, such as survey-based methods, where investors' sentiment is directly measured through regular surveys. These methods can provide valuable insights into investor attitudes and expectations that might be more hidden and not always apparent in publicly available data (Gompers et al., 2008).

[18] These are the areas, as we will show, where this thesis will focus.
[19] Sentiment polarity refers to how polarised (positive or negative) an opinion is.

Modelling market sentiment in the VC industry has several key applications, particularly in the pre-investment (deal sourcing), deal structuring (valuation), and post-investment (portfolio management) phases.

With reference to sourcing and understanding market dynamics, machine learning algorithms can be applied to social media data or startup-related news to forecast investor sentiment and identify emerging trends (Rizk, Mabrouk, & Otrok, 2019).

During the deal sourcing process, sentiment models can help investors in identifying promising startups early by tracking the buzz around certain companies, industries, or technologies. For instance, having the possibility to visualise the sentiment around the new products launched in the market by looking at the reviews it has, can be extremely useful to identify potential investments. This phenomenon is not only relative to products, but might affect entire industries, as Research by Da Rin, Hellmann, and Puri (2013) suggest. In their research they find that VCs are more likely to invest in sectors associated with a positive sentiment, as these areas are perceived to have higher growth potential.

Another critical area where sentiment modelling can be applied is venture valuation. Positive sentiment can inflate valuations, as high investor enthusiasm often leads to increased demand (often hype) for investments in startups. Conversely, negative sentiment can lower the valuations, making it much harder for startups to raise funds. Modelling sentiment allows venture investors to adjust their valuation models to account for these dynamics, potentially leading to more accurate and fair valuations that might mean winning a deal, if valuation is one of the determining factors (Nanda & Rhodes-Kropf, 2013).

Finally, if we consider the area of portfolio management, venture capitalists could use sentiment analysis to monitor their investments and adjust their strategies accordingly. For example, if sentiment around a particular portfolio company or industry shifts from negative to positive, VCs might decide to look for an exit from their investment to leverage the higher valuations that the market might offer in the near future (since sentiment polarity presumably predicts market prices, as previous research found). This proactive approach can help VCs manage risks more effectively and capitalise on positive market trends (Hirshleifer & Shumway, 2003).

Ultimately, the possibility to model and interpret market sentiment offers venture investors a powerful tool for enhancing their investment strategies. By incorporating sentiment considerations into their decision-making processes, VCs can leverage a better understanding of market dynamics, potentially leading to more informed and less biased decisions.

2.2.3. Criticises in Modelling Market Sentiment

Despite the enormous potential, modelling market sentiment in the venture capital space is extremely difficult. There are many challenges, and one of the primary issues is the quality and availability of the data. In fact, unlike public markets where data is relatively abundant and standardised, the VC industry focuses on illiquid, private equity investments that do not generate an observation only when a financing round is announced, or an exit is made. In this setting, sentiment data, especially from social media, can be noisy and may require extensive cleaning and preprocessing to be useful (Schumaker & Chen, 2009).

Another critical aspect is represented by the subjectivity of sentiment analysis. While machine learning models and textual analysis tools have improved, they still struggle with context, sarcasm, and other nuances in human language. This can lead to inaccuracies in sentiment scores, potentially misleading investors (Rizk et al., 2019) and difficulties for NLP algorithms (Pang & Lee, 2008). Moreover, the previously cited study by Loughran and McDonald (2011) cannot be of any help, as the jargon is quite different in the VC asset class.

In any case, the biggest challenge posed by sentiment modelling remains the high noise and ambiguity inherent in sentiment data. Text data, particularly from social media, can be very noisy, with a mixture of relevant and irrelevant information. To avoid this, one needs to select only the parts of the text that are strictly relevant to the analysis, trying to avoid overfitting and ensuring that models are robust enough and can be generalised well across different market conditions (Nassirtoussi et al., 2014).

Moreover, one should not fall in the trap of over reliance on sentiment data, which represents another concern. While sentiment is a useful indicator, it should not be the sole basis for investment decisions. VCs must balance sentiment analysis with fundamental

analysis and other traditional investment criteria to avoid making overly speculative or sentiment-driven investments (Brown et al., 2012). This is in line with the previously mentioned model of hybrid intelligence proposed by Dellermann et al. (2021).

Finally, data quality and availability also pose challenges. While large volumes of sentiment data observations are available, ensuring the quality and representativeness of this data is of utmost importance. Poor-quality data can lead to incorrect sentiment scores and misleading predictions (Schumaker & Chen, 2009). Moreover, some sentiment data sources may not be available in real-time or may require expensive subscriptions, limiting their accessibility to smaller investors. As mentioned earlier, in the last few years all social media platforms have closed their APIs and therefore there is no simple way to collect data from them anymore, aside from scraping (which can be quite hard, again, due to the measures put into place against it).

3. De-Biasing Venture Investment Decisions

As seen in the previous chapter, the venture capital industry is intrinsically characterised by conditions of extreme uncertainty, whereby investment decisions need to be quick and often only based on a handful of information. This often leads to poor decision making and to judgements that are heavily influenced by "gut" feelings (Huang & Pearce 2015), which makes them vulnerable to irrationality and biases, typically associated with what Kahneman (2011) would define as “System 1” thinking[20]. Among all these biases, the herding bias, whereby investors tend to copy the opinions of others, rather than relying solely on their own (Prosad et al., 2015), has been found to be particularly prominent in the venture capital space and can significantly impact the quality of venture investments (Zhang et al., 2021).

In this closing chapter, we will develop a model aimed at investigating which effect does herding have on the quality of venture investments, and in doing so, we develop an algorithm[21] that is able to detect herding in the market from a sentiment analysis of YouTube videos. This can be used to conduct further research, opening a new source of alternative data still unexplored in the financial field, or to de-bias investor’s judgement.

3.1. The Current State of the Research

As described by Gartner’s *Hype Cycle* framework (Gartner, 2024), especially in the initial stages of a technological breakthrough, there is often an overinflated expectation regarding the technology's potential, driven by the crowd's over-enthusiasm. Then, as the technology matures, these expectations tend to vanish, and a more rational and realistic perception of its utility tends to emerge. This cyclical pattern of sentiment aligns very well not only with overconfidence (from which the name “Hype”), but also with herding behaviours, whereby the initial widely shared overconfidence is followed by a correction as the market gains a clearer understanding of the technology's actual value.

[20] In Kahneman’s work, System 1 represents our instinct whereas System 2 our rationality. Often these two systems are in contraposition, and the first one, although quicker and more efficient, is frequently the cause of fallacies in human’s reasoning, namely biases (also known as *heuristics*).
[21] Defined as a set of steps that a machine needs to follow to complete a task.

As previously outlined, to study this phenomenon sentiment analysis emerges as a promising tool. By analysing data from news, blogs, or social medias, it is possible to capture sentiment and a broader range of informational cues that have historically been overlooked[22].

The Research Gap

In first two chapters of this thesis, we have thoroughly reviewed the literature around these themes, and some gaps in the literature have clearly emerged. As shown graphically by Figure 1, most of the research up until now has focused on analysing separately three areas that, however, are strictly related to each other: Venture Capital, the Herding Bias, and Social Media Sentiment. The intersection of these three areas has never been explored, and this thesis will draw from existing literature to explore this space. More specifically, we will *project* the results found in adjacent areas to build our *Herding Index (HI)*[23].

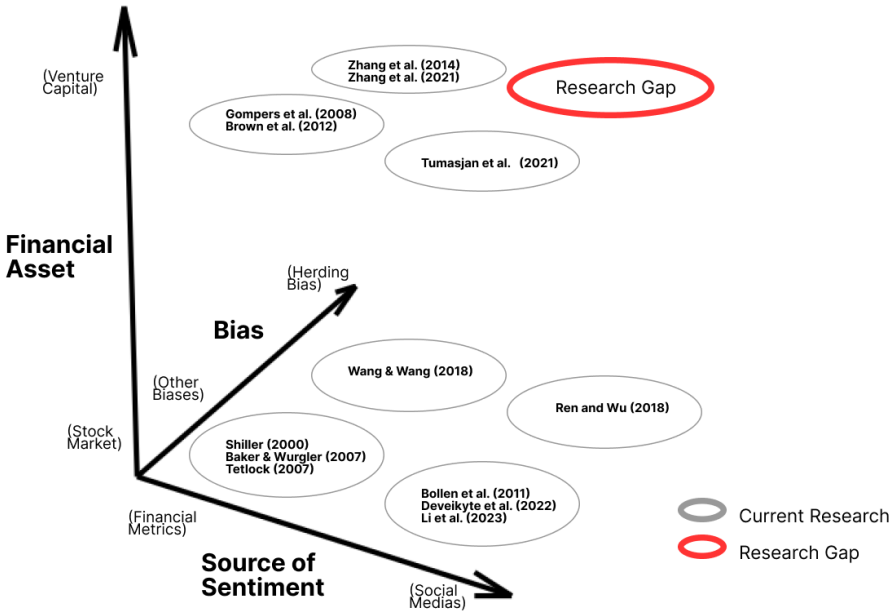


Figure 1: Graphical Representation of the Current Research Gap.

[22] See 2.2 Investor Sentiment for a complete review.
[23] See 3.3.2 Derivation of Sentiment and of *Herding Index (HI)*.

Other Literature Requests

Aside from this research gap, clearly identified also by existing literature (Bloh et al., 2022), this thesis also answers to the direct requests of very recent paper for new, alternative sources of data.

Bain & Zhao (2021) find that a lack of comprehensive data is the primary contributor to poor decision-making in the venture capital sector and a more complete view of sentiment could help the decision-making processes. In fact, current research has little explored new sources of data, as they are not easily retrievable.

Although it is true that, as pointed our earlier in the previous section[24], the academic world seems to have moved some steps in this realm by accessing data from Twitter (today the main source of sentiment data), the changing data landscape and the harshening conditions on the ease of access, created an environment in which it has become extremely difficult to conduct such studies[25].

In this setting, Vanacker et al. (2019) and Tumasjan et al. (2021) openly call for access to new data sources that could yield a more nuanced understanding of market dynamics, offering new avenues for research and potentially more accurate predictive models.

The purpose of this thesis, apart from filling the current research gap, is strictly bound to this necessity.

A New Measure of Success for Venture Investments

Finally, in light of the need to measure the overall quality of a venture investment, and not only its short-term return, we will also create a new dependant variable that is able to really capture the idea of *success* in a venture investment (Cao et al., 2022).

3.2. Hypothesis Formulation

The aim of this research should be clear by now: create a tool to access new sources of sentiment data and link these to the herding behaviours in the venture capital space.

[24] See 2. The Use of Alternative Data in Investment Decisions.
[25] Even Twitter (now X), after Musk's acquisition, decided to close its API to the public (TechCrunch, 2024).

Having built the model[26], as a first step, we will investigate the relationship that exists, if any, between herding and the success of investments in the venture capital space. From the previous literature, we infer that these two factors should be correlated.

> *Hypothesis 1 (HP0[1]): The herding bias does not significantly influence the quality of venture investments decisions*.

Further, as previously stated, the venture capital market seems to be bound to *irrational exuberance* and *hype cycles* which are commonly thought to be a consequence of major technological shifts. However, in the literature, the herding bias has never been explicitly linked to these. Instead, it has always been studied as a phenomenon characterizing indiscriminately the whole venture capital industry. Thus, we want to test whether technological shifts are associated more prominently with the herding bias compared to the rest of the industry.

> *Hypothesis 2 (HP0[2]): The herding bias is not strictly bound to major technological shifts.*

As already mentioned, to test these hypotheses, we will first need to build a model that is able to derive herding from sentiment, and then to link it to performance of venture investments. We will reject the null hypothesis only with a confidence level of at least 99% (1% significance level α).

3.3. Methodology

To build the model, the first step was to develop an algorithm that could extract information on venture investor's sentiment at each point in *time* and for a certain *trend* leveraging alternative sources of data, namely YouTube videos. From this data, a proxy of how much herding is present in the market is derived and then put into relation with the quality (or success)[27] of venture investments realized in those years.

3.3.1. YouTube Sentiment Analyser

The first step to develop the model, was to find the data necessary to run the analysis and test our hypotheses.

[26] Which we will do in chapter 3.3 Building the Model.
[27] Quality and success of venture investments will be used interchangeably for the rest of the thesis.

YouTube as a Source of Data

As previously mentioned, up until now most of the literature has focused on examining sentiment derived from *semi-structured* [28]sources like news articles, tweets, blog posts etc.

In this thesis, acknowledging the above-mentioned literature's desire to advance the analysis to new, alternative sources of data, the recent restrictions and closures that affected the access of Social Medias APIs, and the strengthening defence mechanisms put into place against scraping algorithms[29], the choice was made to tap into a new and completely unexplored area, that is the last available space that can really capture what is going on through investor's minds: YouTube.

Aside from these "necessary causes," the reasons behind the choice of YouTube as the primary source for investigating venture capital investors' sentiments are mainly two. First, this platform offers a unique blend of in-depth content from tech experts and direct engagement of investors, which are two key elements that are not typically found in news articles or blogs[30].

In fact, unlike traditional news articles, that are often written by a small group of journalists and might be subject to editorial biases, YouTube allows for a much broader range of voices, including industry experts, journalists, people from academia, actual investors and founders who share their perspectives and insights through video content. Moreover, videos on YouTube often are longer and delve deeper into topics, allowing for more comprehensive explanations and analyses compared to the often brief and surface-level coverage found in blogs or news articles. This depth, combined with the platform's wide reach and the ability to capture both the creator's sentiment and (eventually) the audience's reactions, makes YouTube a tool particularly fit for gauging the sentiment of investors.

[28] Data that does not have a tabular structure.
[29] TechCrunch (2024).
[30] This is particularly true in the periods of analysis (2010-2020), when YouTube was very popular.

Scraping YouTube Transcripts

Having a truthful representation of venture investor's sentiment is not exactly straightforward. As seen in the previous sections, social medias are found to be a good proxy of it and thus worth modelling.

YouTube is much harder to work with for two main reasons: it is hard to access, as one needs to download and analyse often messy transcripts of conversations, and it is hard to model because often data is a very noisy.

Notwithstanding these difficulties, I wrote an algorithm in python that, given as input the YouTube channels, the topics, and the years of analysis, calculates what is the general sentiment around the chosen *trend* and its evolution over time. This is perhaps the most important element of this thesis, as it is a first of its kind tool, and enabled the development of this whole thesis. For further details on this, see Appendix 3.

The Choice of the YouTube Channels

To select the YouTube channels to analyse, several criteria were taken into consideration.

First of all, in order to collect enough sentiment observations, we are going to consider only channels that are at least 14 years old, so that there is available data starting from 2010, the starting point of the analysis. Then ending point of the analysis will be 2021[31].

Moreover, the aim was that of grouping various kinds of information channels: the more typical and formal journalism (news) and the more informal, and perhaps insightful, interviews of investors and tech experts (the tech conferences and the podcasts). With these two quite different types of data, one should be able to have a complete overview of what is the venture investor's sentiment around industries, technologies, or specific investments. Moreover, this choice allows the model to not analyse only news, where catchy headlines might not reflect the correct sentiment, or only conferences, where views and opinions might be expressed more mildly, but a good mixture of the two.

[31] The starting point, 2010, is an obliged choice as the library used to retrieve transcripts of YouTube videos could not go further back in time. The ending point, 2021, has been chosen balancing the need to give the youngest startups in the sample enough time to grow (from 3 to 4 years approximately) while at the same time collecting the most amount of data as possible.

In the end, the identified sources were The Verge (news), Bloomberg (news), CNBC (news), TechCrunch (news and podcasts), Week in Startups (podcast)[32].

Inside these channels, the videos analysed were all those in the period 2010 – 2021, without any discrimination, as the filtering of the information will be determined by which *keywords* are contained in their transcripts, as explained in Appendix 3.

Choice of Trending Topics

The *trends* analysed are some of the technologies or industries that historically characterised the venture capital space through the years 2010 – 2021 (Startup Daily, Gartner, 2024), namely:

- Autonomous Vehicles (Technology)
- Blockchain (Technology)
- Clean Tech (Industry)
- Cloud Computing (Technology)
- Ecommerce (Industry)
- Health Tech (Industry)
- Consumer Tech and IoT (Industry)
- Mobile Apps and Social Media (Technology)
- Quantum Computing (Technology)
- Voice Assistants (Technology)
- Virtual and Augmented Reality (Technology)

The choice of analysing a mixture of industries and technologies comes from the desire to understand in which of the two the herding phenomenon tends to manifest more if that is the case (testing H2). Moreover, modelling could differ between the two based on how people refer to them in a sentence.

3.3.2. Derivation of Sentiment and of *Herding Index (HI)*

Once the tool to download and filter the information present in YouTube transcripts has been created, it necessary to analyse the relevant sentences and extract the relevant sentiment scores from each of them.

[32] Consulted articles were Sourcescrub (2023), The Muse (2020).

This is achieved by using an NLP (Natural Language Processing) library called "TextBlob"[33], with which we can calculate, for each sentence stored, two scores:

1. The sentiment polarity score: this indicates how positive (or negative) is the sentiment in the sentence. The sentiment polarity score ranges from -1, representative of a negative statement, to 1, representative of a positive statement. As an example, "The sun is such an amazing star" scores 0.30 whereas "The last scene of the movie was quite unsettling" scores -0.25.
2. The sentiment subjectivity score: this indicates how objective (or subjective) each sentence is. The sentiment subjectivity score goes from 0, for an objective statement, to 1, for a very subjective statement. As an example, "I believe iPhone is the best phone currently on the market" scores 0.35, whereas "iPhone sales have doubled from last year" scores 0.07.

Additional to these two scores, for each sentiment observation we also collect the date of the observation, and the number of observations made in that date, so that it will be then easier to reconciliate this data with the market's one.

Combined Sentiment Score (CSS)

Having collected all this data, a *Combined Sentiment Score (CSS)* is then calculated for each sentiment observation as the product of *Sentiment Polarity* (how positive or negative each sentence is) and *Sentiment Subjectivity* (how much a sentence embeds an opinion or a fact).

$$Combined\ Sentiment\ Score\ (CSS) = Sentiment\ Polarity\ x\ Sentiment\ Subjectivity$$

As mentioned previously, the *Sentiment Polarity* score goes from -1, representative of a negative statement, to 1, representative of a positive statement and the *Sentiment Subjectivity* score goes from 0, for an objective statement, to 1, for a very subjective statement. Thus, the *Combined Sentiment Score (CSS)* will range from -1 to +1 and combine the two effects in a single score that embodies both sentiment polarity and sentiment subjectivity.

[33] Despite being quite simple compared to models like BERT, TextBlob allows for a more efficient and quick analysis of large datasets, with similar level of confidence.

The reason for the calculation of this score is that the subject of our analysis are opinions, and not facts. Although it is true that facts can often be brought up to support an opinion, in general, somewhere in the discussion, there will always be an explicit reference to the subjective opinion of the person, and this is what we want to capture to correctly model sentiment.

Having developed this *Combined Sentiment Score (CSS)* we can properly capture sentiment in the YouTube transcripts and can move on assessing whether this can be somehow linked to herding tendencies in the market.

Derivation of the Herding Index (HI)

To move forward with the analysis, we first need to review three key findings in adjacent areas, already mentioned in the previous sections, which set the theoretical bases for the derivation of a metric that can capture herding in venture investor's opinions.

First, Ren and Wu (2018) found that the volatility of social media sentiment (convergence or divergence of sentiment) can be a good proxy of herding in the stock market. This means that, generally speaking, a low volatility in the sentiment values tends to be associated with herding in financial markets, whereas a higher volatility tends to be associated with the opposite (low herding). This logically makes sense, because if all agents in the market have the same opinion, and the volatility of the sentiment is low, then there is a higher chance that they are not thinking independently.

Assuming this result also holds true for the venture capital market[34], we can state:

$$\exists \rho_{\sigma(CSS);\,\text{Herding}} > 0 \quad s.t \quad \left|\rho_{\sigma(CSS);\,\text{Herding}}\right| > \rho_{crit} \quad and \quad p < \alpha \qquad (1)$$

Secondly, numerous papers (Gompers & Lerner, 2001; Wang & Wang, 2018; Zhang et al., 2021) find that the "herding attitude" is mostly present in investors in the context of extreme market conditions, namely during booms and crashes.

Considering that, following literature's findings (Tetlock, 2007; Bollen et al., 2011; Nassirtoussi et al., 2014), sentiment should be strongly correlated to the realizes returns in the market, we can state that whenever the *Combined Sentiment Score (CSS)* is very

[34] The first assumption of our model is that the stock market and the venture capital market are related to investor sentiment in similar ways.

high or very low, the herding phenomenon should prevail in the market, and therefore (1) should be even more evident[35].

Assuming again that the stock market and the venture capital market have similar relationships with investor sentiment, then we can state that:

$$\sigma(CSS) \to 0 \quad as \quad CSS \to +1^{-} \quad or \quad CSS \to -1^{+} \qquad (2)$$

See the figure below for a clearer visualization of assumptions behind the derivation of the *Herding Index (HI)* starting from *the Combined Sentiment Score (CSS).*

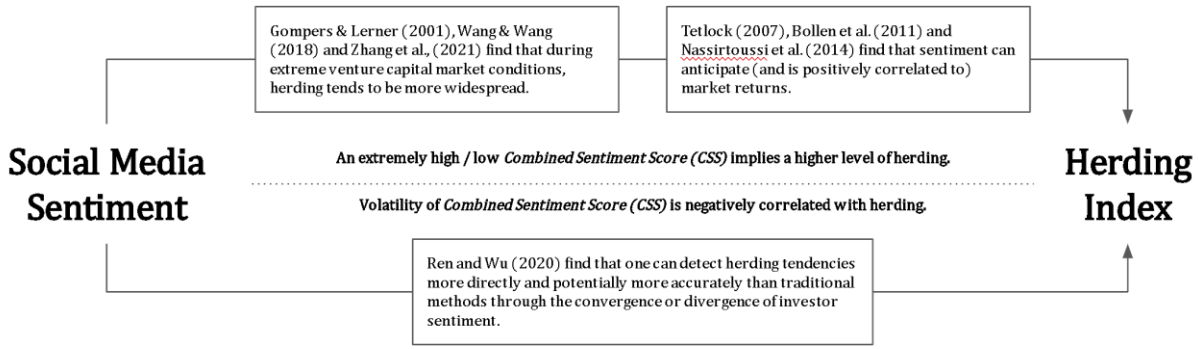


Figure 2: Derivation of the *Herding Index (HI)*.

To build the *Herding Index (HI)* we leverage (1) and (2) to create a unique score that can proxy the amount of herding in the investor's sentiment. More specifically, using (1) I first derived for each year of observation of the *Combined Sentiment Score (CSS)*, its annual sample standard deviation. Then, according to (2) we can state that the volatility of the *Combined Sentiment Score (CSS)* represents a good proxy of herding as long as the market, and thus the *Combined Sentiment Score (CSS)* itself in our model, presents extreme values.

We are now ready to define the *Herding Index (HI)* as:

$$HI_t = \begin{cases} \sigma(CSS_t), & CSS_t < Q1(CSS) \ \ or \ \ CSS_t > Q3(CSS) \\ \nexists, & Q1(CSS) \geq CSS_t \geq Q3(CSS) \end{cases} \qquad (3)$$

[35] It is essential to proxy market returns in extreme conditions with the *Combined Sentiment Score (CSS)* for two reasons: returns are often hard to observe in real time in an illiquid market such as the venture capital one and, even in the case in which they were, sentiment has been found to anticipate them, and this grants the model a predictor power for future returns, which is extremely valuable if any investor want to use it as a tool to de-bias its decision making.

Whereby:

- *HI* is the *Herding Index*.
- *CSS* is the *Combined Sentiment Score.*
- *t* is time (measured in years).
- σ() indicates the sample standard deviation.
- μ () indicates the sample average.

Note how the score does not exist in the middle 50% of the CSS observations, as this is not the area in which we are interested in, as empirically from (2) we know it is not of our interest. Effectively, this will result, when running OLS regressions, in quartile regressions.

Finally, note how the only assumption we had to make to create this score, thanks to the already developed literature analysed in the previous sections, is that stock markets and venture capital markets are related to investor sentiments in similar ways.

3.3.3. The Final Sentiment Dataset

The final dataset containing observations of the *Combined Sentiment Score (CSS)*, in the above-mentioned years and *trends*, was derived by running the algorithm over a total of more than 80 hours of videos and contained a total of 30,541 observations.

Overall, the distribution of sentiment seems to be quite centred around its mean of 0.08 with a slight skew to the right (Figure 3a).

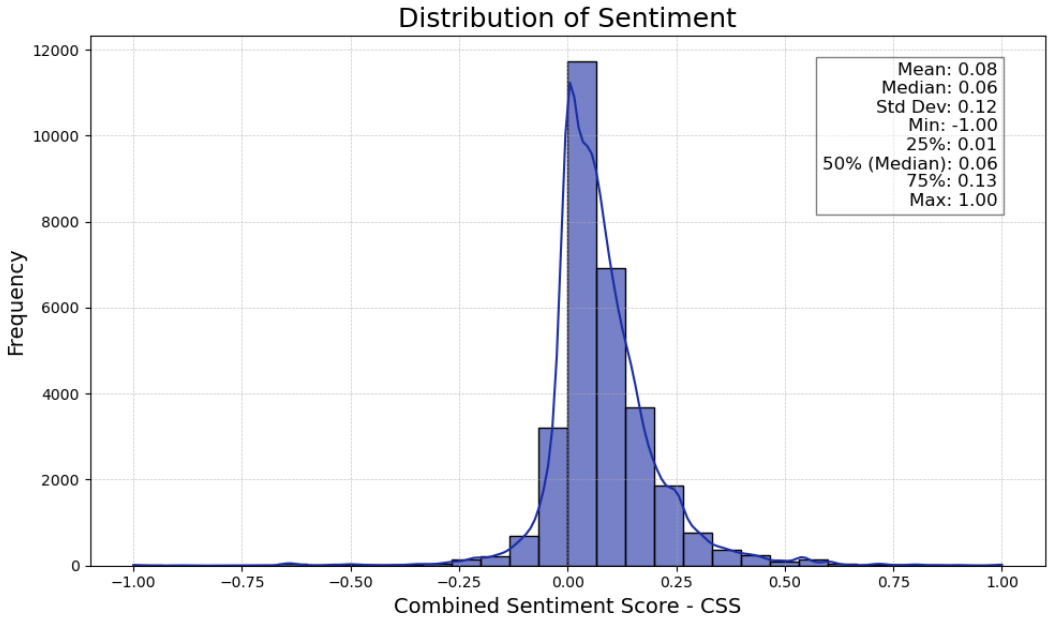


Figure 3a: Histogram and Kernel Density Estimation of the *Combined Sentiment Score (CSS).*

Moreover, although the dataset is well distributed over time (see Figure 3b) but not as much across *trends* (see Figure 3c). Unfortunately, this only depends on how many times these topics were discussed in the analysed videos, and therefore cannot be controlled for. As we will have to estimate parameters, it is important to have a sufficient sample size each year for each *trend.* For this reason, we will exclude "Quantum Computing" and "Voice Assistants," respectively with 179 and 55 observations. After this change, the sample now counts a total of 30,220 observations and its key statistics and distribution remain basically unchanged.

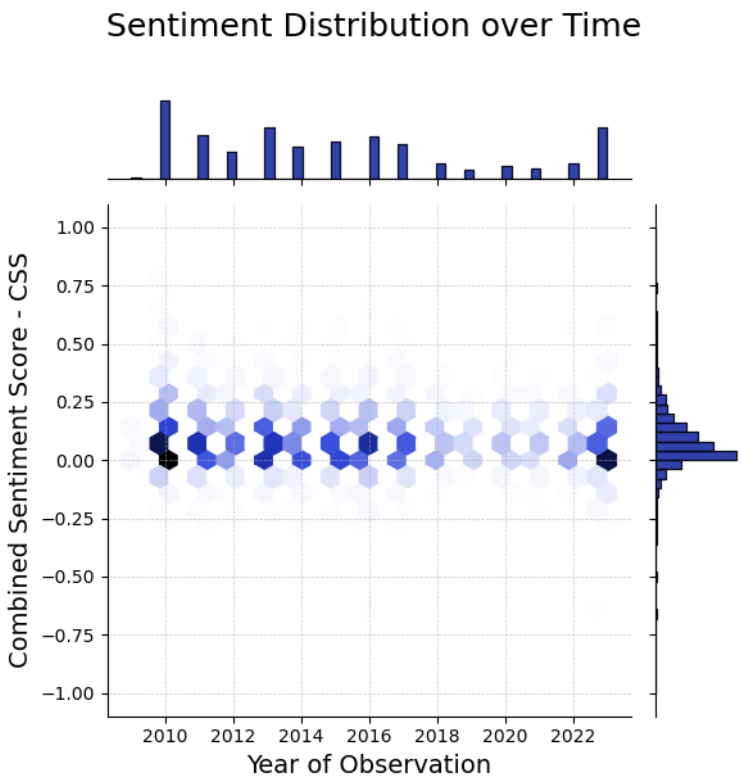


Figure 3b: Sentiment over Time.

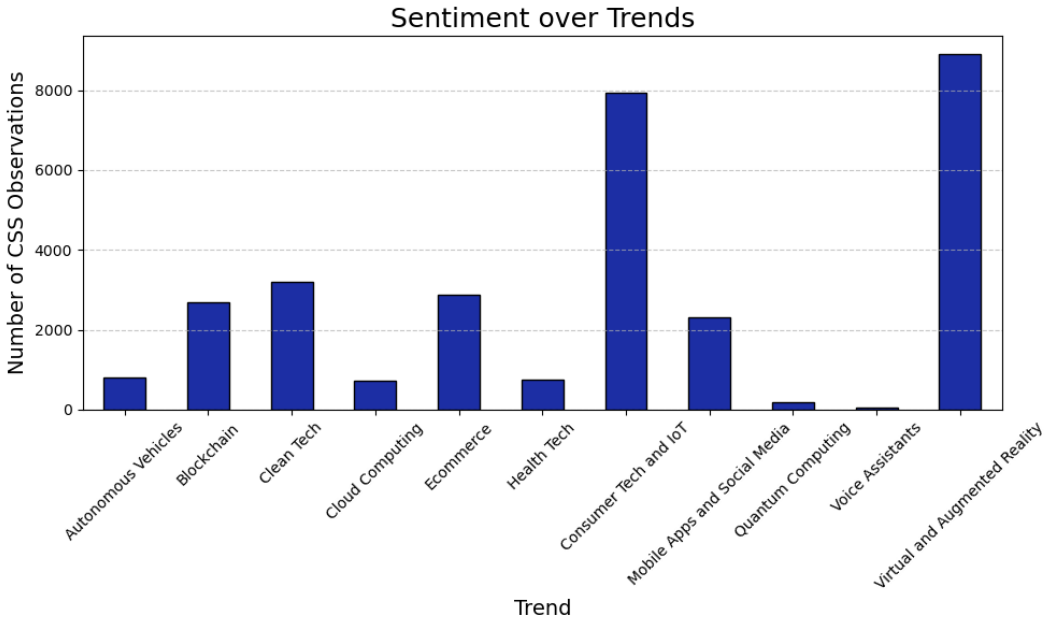


Figure 3c: Sentiment across Trends.

This dataset is then divided into two parts: a training dataset and a test dataset, useful to understand if the modelling can find applications outside its training, particularly important for (most likely) very noisy data like this, where the risk of overfitting might be high. As literature suggest (Cao et al., 2022), the model should be trained using 70% of

the data, while 30% of it was kept for the tests. This is the approach we will follow; further details are given in the next section.

Deriving the Herding Index (HI)

At this point we derive the *Herding Index (HI)* using formula (3). To do so, we first normalize CSS scores for each trend, so that they can then be aggregated and compared[36].

To test our hypotheses, we need to better understand what kind of relation exists, if any, between herding and venture capital returns.

To do so, we will first define a new metric that can capture the success of a venture investment. Then, we will use a linear model to examine data more in depth. This choice is made to avoid over-fitting a dataset that, most likely, is very noisy and to avoid adding complexities in an analysis that is already full of caveats.

As previously mentioned, to provide robustness to this analysis, we will split the sample in two parts in order to train it and then test it.

3.4. A New Definition of Success in Venture Investments

3.4.1. Measuring Success

Several criticises can be moved to how, when training a model or testing some hypothesis on a certain dataset, the success of venture investment is measured.

Typically, most of the literature has always focused on IPOs and M&As[37] as primary metrics of success in venture investments, and while they provide an immediate financial return for investors, these metrics fail to capture the long-term viability and health of the startup (Gompers, 2001; Deias & Magrini, 2023).

Ultimately, while it is true that the short-term outcome is what matters most to early-stage investors, this will not be the focus of this thesis. As a matter of fact, in this piece of research, we are interested in the overall success of a venture investment, and not simply in the returns realized by the first investor that enters with an equity investment.

[36] This is an essential step, as otherwise we would risk considering only one of the trends that reached more extreme *CSS* values than the others.
[37] Initial Public Offerings, Mergers & Acquisitions.

Arguably, a good venture investment does not manifest itself necessarily in a short time frame but might even take more than ten years[38].

As Cao et al. (2022) emphasise, the dependent variable should reflect the industry-specific definition of "success", which should be tailored to the environment. In the venture capital industry, “success” can mean many different things, and can be measured in as many ways, from typical financial metrics to alternative ones, like investor’s quality or network (Zava & Caselli, 2023; Zava, 2024).

However, in the last few years, the venture capital space has evolved to become a vastly different asset class distancing itself from the more typical private equity industry. In this setting, fund returns are not driven anymore by a generally satisfactory performance in its investments, which can be proxied by the number of exits, as the literature typically does. Instead, the modern definition of “success” increasingly centres around having the ability to generate the so-called *home runs*—investments that return at least 3x or more and that often constitute the great part of a fund’s returns. As a percentage of the invested capital, these investments should be at least 5% of invested capital (Oswald, 2021).

Following this line of thought, one could argue that a market for an industry or for a particular technology is performing well if the top 5% of companies in that sector are achieving significant returns[39]. This approach would also align well with standard industry practices, whereby evaluating the performance of the top quantiles is a common metric for assessing both the startup market and VC fund performance. This coherence with current market practices ensures that the analysis we are going to do stays relevant and grounded in the realities of the venture capital space.

[38] More than the maximum time that the company can remain in the portfolio of the typical venture capital firm.

[39] This seems reasonable if we assume that early-stage financing, in its amounts, is quite similar between companies and across industries. This would mean that the top 5% of the companies in the market (with respect to valuation) will have received initial financing no different than any other, and therefore will correspond to the 5% of the invested capital by the market. This is especially true for very early-stage investors.

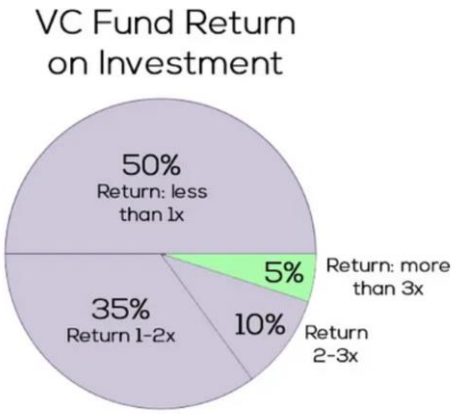


Figure 4: Typical VC Portfolio Returns.

For these reasons, the dependant variable used was the average valuation of the top 5% of the companies founded in that year.

It is an extremely innovative metric, not easy to derive from the initial dataset, but that perfectly embeds the essence of "success" in the venture capital industry. Moreover, it implies no loss of information compared to using as a measure of success the number of IPO / M&A exits made by VCs. Finally, apart from a few outliers[40], arguably most of the startups tend to be valued using the VC method[41] in the initial stages of life, and therefore tend to receive the same amount of investment for approximately the same percentage of equity.

3.4.2. The Final Valuations Dataset

To understand if the *Combined Sentiment Score (CSS)* derived from the algorithm made sense, I collected a dataset of venture capital investments made between 2010 – 2021 in the same *trends* previously mentioned.

To achieve this, I used the CB insights' database. Looking at the *Deals* section, I applied four filters.

- I only considered the period 2010 – 2020.
- I filtered for only VC backed startups.

[40] A category of outliers is represented by foundation models, as for example the recent 1bn funding round obtained by Safe Superintelligence Inc. (https://www.theverge.com/2024/9/4/24235957/openai-cofounder-raises-1-billion-for-his-openai-rival).
[41] VC method (2007): https://vcmethod.com/.

- I selected only Europe and the United States as geographies[42].
- I filtered the results for startups that were containing, in the description of their business, the keywords that indicated that they were part of a certain "trend"[43].

The final dataset is composed of 21,666 deals[44] which give us a total of 7,140 companies that were financed in that period, and for which we have a *latest valuation* value, independently from whether they were successful or not[45].

For each year and *trend,* we then calculate the average top 5% latest valuation for that year and *trend*, following the logics explained earlier, and are left with a sample of 150 observations.

Plotting this data, we see that even when considering only the greatest startups a fund has financed, it is rare to see strong returns. Nonetheless, we can see how it resembles what we mentioned earlier about performance in the market (Figure 5a). The data clearly shows how most of the companies do not reach extraordinary valuations, but there are some "bumps" in the distribution that indicate the *home runs* to which we were referring to previously. In today's venture environment, only a handful of companies end up performing very well, which typically means reaching a valuation higher than 1bn dollars[46].

[42] Because also the YouTube channels selected for the sentiment analysis are strictly pertinent to these two geographies.
[43] For a complete list of these keywords, please see the appendix.
[44] In order of listing in the previous section, numbers of deals for each *trend* are 610, 3187, 3713, 1668, 412, 3247, 2885, 3140, 766, 2038.
[45] For our definition of "success," we want to avoid a *selection bias* in selecting only successful investments, but consider all, even failed, startups that were funded. After all, they are also part of the VC portfolio as examined in the previous section.
[46] That is, becoming a unicorn.

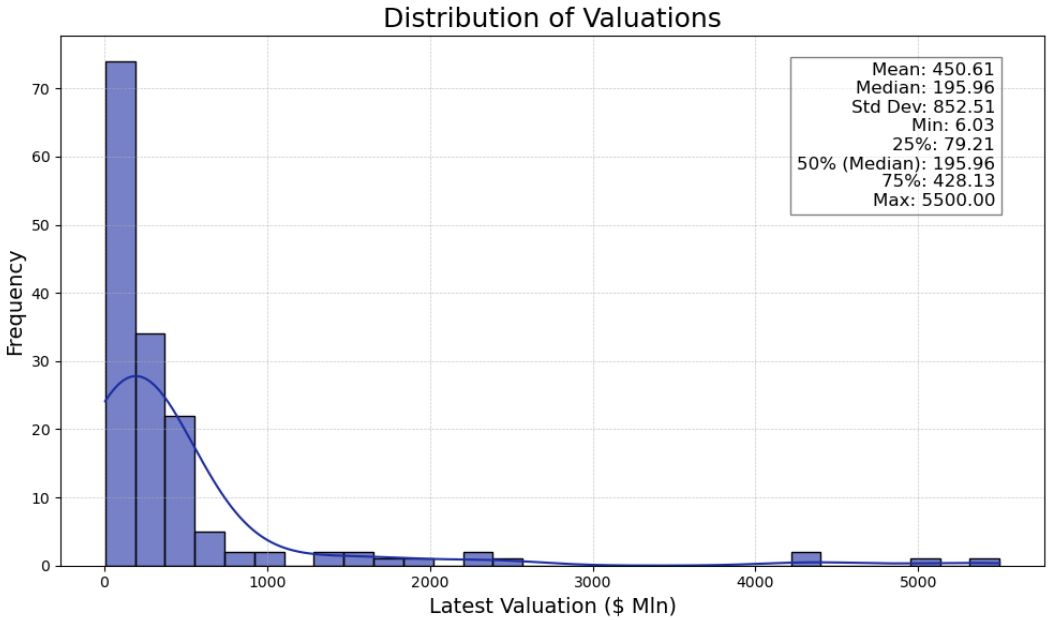


Figure 5a: Histogram and Kernel Density Estimation of Valuations.

Analysing better this data (Figure 5b), we can notice how valuations seem to be uniformly distributed over time, and especially around the years 2010 – 2020 we can see that startups should have become mature enough to reach a valuation that is more "stable" and therefore can be used as a good proxy of success[47].

Looking at the distribution of data across *trends* (Figure 5c), we notice how, apart from "Voice Assistants" and "Quantum computing" which have already been excluded from the analysis, there is significant lack of data for the "Autonomous Vehicles" *trend*, and thus, this was also excluded from the analysis[48].

[47] In fact, the average startup takes approximately 5.5 years to become a unicorn (Orios Venture Partners, 2023) and thus, since the sample considers 2020 as the last year of observation, all the startups in it should have reached (or almost reached) maturity, and therefore a valuation that reflect their value and potential (Economic Times, 2024).

[48] Having less than 200 observations over the time period, would mean having less than 20 per year, and therefore the top 5% would be driven only by the best performing company.

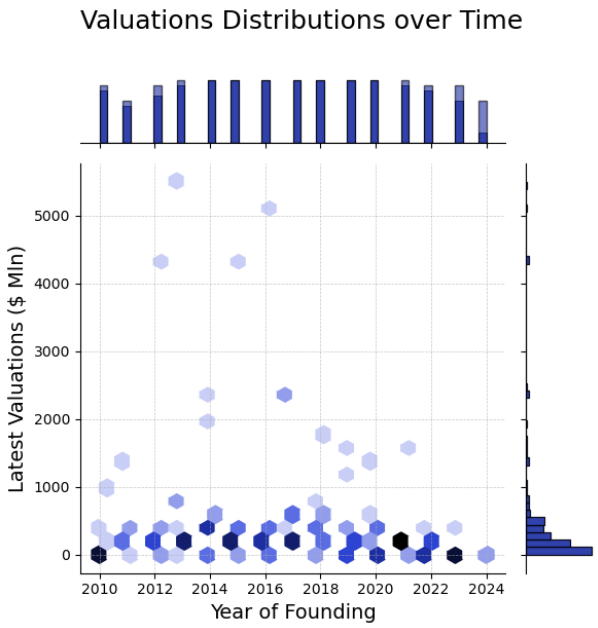


Figure 5b: Valuations over Time.

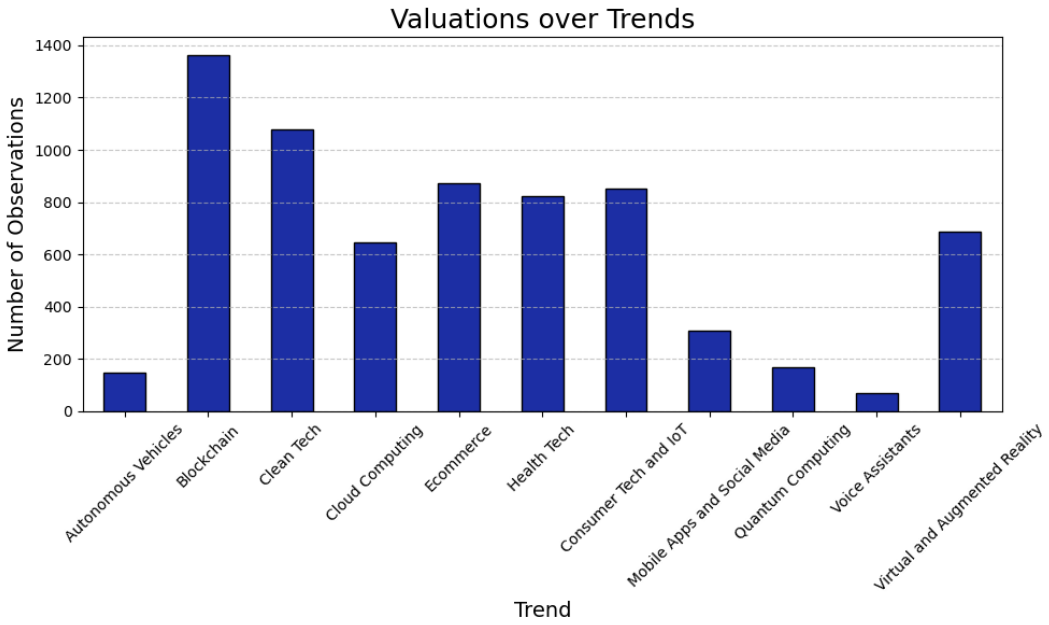


Figure 5c: Valuation Observations across Trends.

3.4.3. A comparison with Previous Metrics

Finally, if we compare this new dependant variable to the more typically used percentage of exits made on the investment (either through IPO or M&A) we see that there is clear correlation between the two, and for the most part they tend to move together (better investments are exited and at a greater valuation). However, sometimes they do not.

It becomes evident how there is a loss if information when the only thing considered is whether an exit has been realized and not at how much. Looking at Figure 6, we show how

for the Heal Tech *trend,* the typical success metric fails two capture two of the most valuable startups in the dataset, namely Capital Rx (902Mln$) and Brightinsight (701Mln$), both founded in 2017.

In our model, this is an extremely important observation because it is one in which the sentiment polarity is also exceedingly high, meaning it will be counted in the regressions we will do and is part of the *Herding Index* (top quartile of *CSS*).

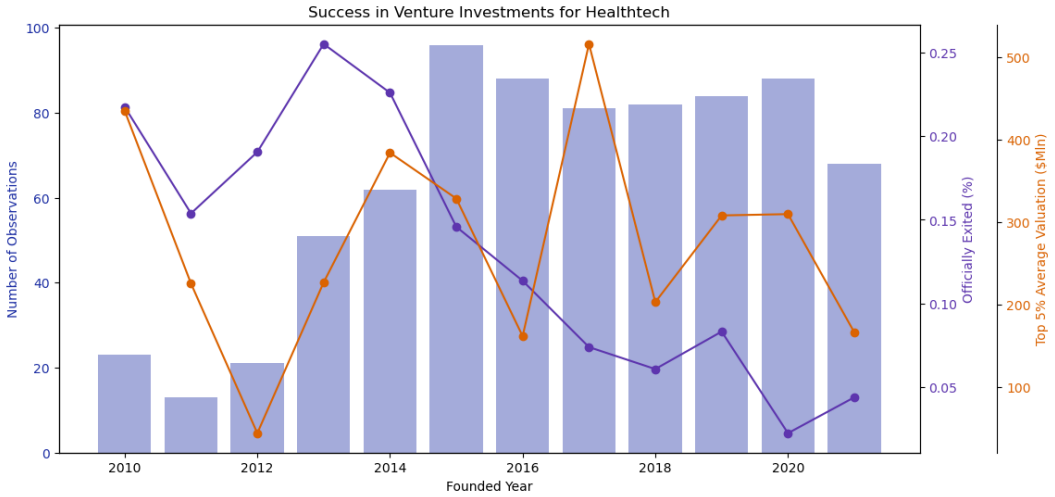


Figure 6: A Comparison of *Success* Metrics in the Venture Capital Industry.

3.5. Analysis of the Results

As the previous literature review explains, typically the methods used to model alternative data and sentiment are two: machine learning (especially ANNs) and linear regressions. Out of these two, I chose to use linear regressions for a few reasons.

First, we expect to see some degree of linearity in our theorized model. In fact, as explained above, what we would expect to see is a direct, positive relationship between the *Herding Index (HI)* and the market results.

Additionally, our sample is not very numerous. As we will specify later, after cleaning the data[49] we have a total of 8 *trends* and of 113 observations.

[49] After the exclusions of the samples that were not numerous enough, see 3.3 and 3.4 for reference.

3.5.1. Preliminary Analysis

First, we investigate very broadly the relationship that exists between the *Combined Sentiment Score (CSS),* its volatility and the performance of venture investments.

We find that the raw *Combined Sentiment Score (CSS)* has no significant relationship on the quality of the investments made (Figure 7a), and this is coherent with the findings of Tumasjan et al. (2021) that found not significant relationship between sentiment and long-term performance of investments.

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		2.282
Methond:	Least Squares			Prob (F-stat):		0.134
N. Observations:	91			Log-Likelihood:		-121.25
R-squared:	0.025			AIC:		246.5
Adj. R-squared:	0.014			BIC:		251.5
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.1273	0.098	1.305	0.195	-0.066	0.321
CSS	0.1769	0.117	1.51	0.134	-0.056	0.41
Durbin Watson:	1.935					
Jarque-Brera (JB):	4.667					
Prob (JB):	0.0969					

Figure 7a: *Combined Sentiment Score (CSS)* and Success of Venture Investments.

However, we also see that there is some sort of degree of positive linearity between the volatility of the *Combined Sentiment Score (CSS)* and the success of venture investments (Figure 7b).

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		2.996
Methond:	Least Squares			Prob (F-stat):		0.087
N. Observations:	91			Log-Likelihood:		-120.9
R-squared:	0.033			AIC:		245.8
Adj. R-squared:	0.022			BIC:		250.8
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.1265	0.097	1.302	0.196	-0.067	0.391
Stdev(CSS)	0.1913	0.111	1.731	0.087	-0.028	0.411
Durbin Watson:	2.001					
Jarque-Brera (JB):	5.109					
Prob (JB):	0.0777					

Figure 7b: Volatility of *Combined Sentiment Score (CSS)* and Success of Venture Investments.

To better investigate this relationship, we will now calculate the *Herding Index (HI)*, which, as previously explained, is the standard deviation of the *Combined Sentiment Score (CSS)* when this is in its bottom or top quartile.

3.5.2. Regression Results

Herding Index (HI) and Market Performance

To have a quick look at the data, we naively plot it (see Appendix 1) and find that there clearly is some kind of relationship between the *Herding Index (HI)* and the market performance. However, this relationship is hard to understand, as many times, the *Herding Index (HI)* seems to anticipate a good market performance, whereas some other times it seems to come after it.

This can be partly explained by the fact that not necessarily the year of founding corresponds to the year in which an opinion about that certain company, industry or technology is formed in the market. For example, public opinion might ignore it for some time or, on the opposite, anticipate it by years. In Figure 8a and Figure 8b, two of the most evident examples of this phenomenon are reported[50].

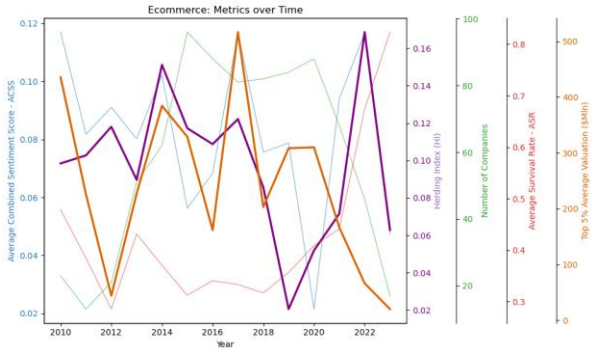


Figure 8a: Blockchain Success and Herding.

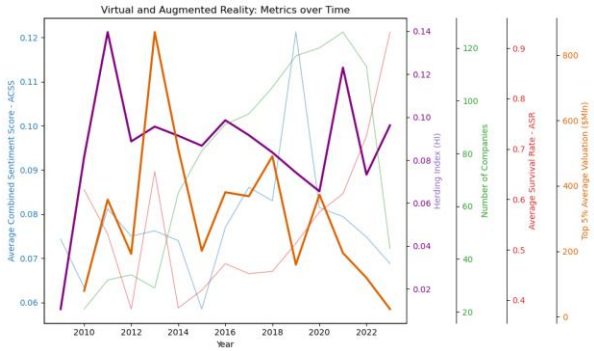


Figure 8b: AR&VR Success and Herding.

Training and Testing Samples

To train and test our model, we divide the sample in two parts which, as previously mentioned, will be 70% for the training of the model and 30% to test its effectiveness.

[50] Remember that the dependant variable that we have developed, to put it in simple words, measures how many successful startups were founded for each year, where *success* is measured as the average valuation of the top 5% most valued startups founded in that year. As an example, if our sample were to include Facebook, then there would be a company, founded in 2004, whose latest valuation is around 500 bn$, and this would have driven up by a great amount the average valuation of the top 5% most valued companies founded in 2004.

- The *training sample* is made up of Blockchain (technology), Cloud Computing (technology), Clean Tech (industry), Ecommerce (industry), Health Tech (industry) IoT & Consumer Tech (industry), Mobile Apps and Social Media (industry), for a total of 20,061 (2,961 + 728 + 3,027 + 2,865 + 739 + 7,930 + 2,351) observations, which correspond to the 70% of the total observations.
- The *test sample* is made up of just the last collected sample, and that is the Virtual and Augmented reality (technology), for a total of 8,909 observations, which correspond to the remaining 30% of the sample.

Summing up the two samples we obtain 29,510 *CSS* observations and not the previously mentioned 30,541. This is because, as mentioned earlier, I decided to exclude three samples, namely the "Quantum Computing", "Voice Assistants" and "Autonomous Vehicles" due to the scarcity of observations[51], which were not numerous enough to have a satisfactory estimate of the volatility of the sentiment (necessary for the derivation of the *Herding Index*)[52].

This leaves us with a total of 8 *trends* and of 113 observations of the *Herding Index (HI)* and of our new dependant variable. Having chosen the period 2010-2021, we also need to exclude observations outside of this range, and we are left with a total of 93 observations[53].

Linear Regressions

At this point, we used the OLS method to regress the standardized[54] investment performance (that is, our newly developed dependant variable) on the standardized *Herding Index (HI).*

For the training sample, I found the results illustrated in Figure 9a and Figure 9b. These show a significant (99.6% confidence level) linear relationship, able to explain a small but relevant part of the variability in the results (R-squared of 0.201, adj.R-squared of 0.180). Moreover, the Jarque-Bera tests indicates that residuals should be normal, and therefore

[51] Respectively, these *trends* contributed for 179, 55 and 797 CSS observations.
[52] 179 observations over 10 years are approximately 18 observations per year, not sufficient to properly estimate the average CSS and the volatility of the CSS. Needless to say, this is even more true in the case in which we only have 55 observations.
[53] For three *trends* we are missing the 2010 observation.
[54] I had to standardize all the observations and the market results so that I could aggregate data between different *trends*.

there is no clear structure in the errors that the model does not capture (normality of the residuals is not rejected).

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		9.584
Methond:	Least Squares			Prob (F-stat):		0.00368
N. Observations:	40			Log-Likelihood:		-47.953
R-squared:	0.201			AIC:		99.91
Adj. R-squared:	0.180			BIC:		103.3
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	-0.0262	0.133	-0.197	0.845	-0.295	0.243
Herding Index (HI)	0.4176	0.135	3.096	0.004	0.145	0.691
Durbin Watson:	2.12					
Jarque-Brera (JB):	2.047					
Prob (JB):	0.359					

Figure 9a: *Herding Index (HI)* and Success of Venture Investments.

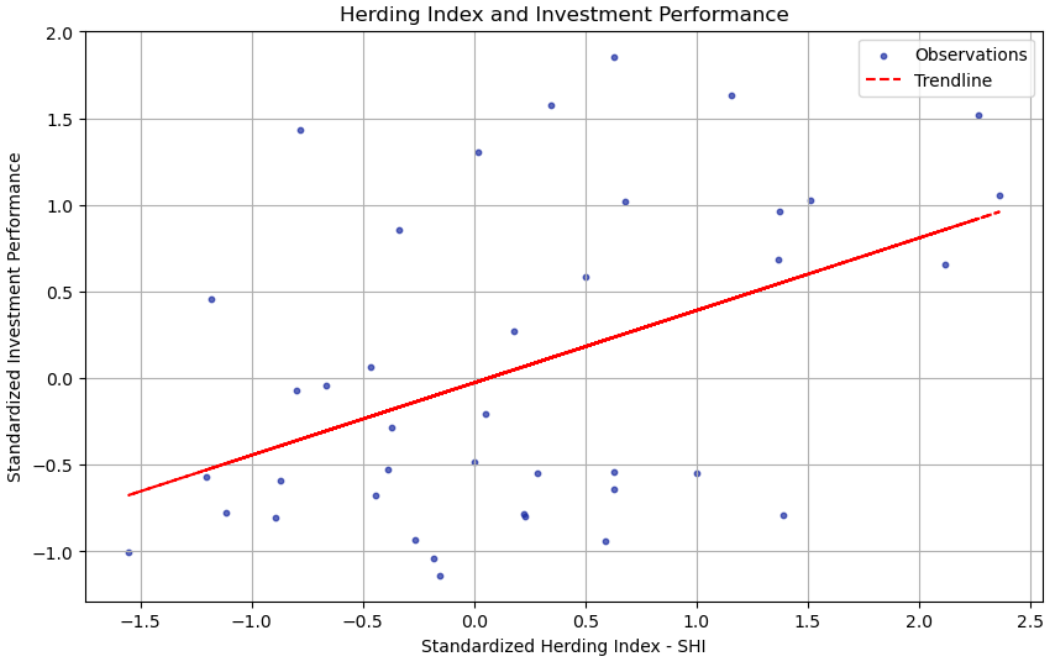


Figure 9b: Graphical Representation of the Model.

These results are coherent with expectations, as the success of investments, even in extreme conditions of the market, is most likely driven by other factors, as outlined in the previous sections[55] such as the quality of the founding team, the size of the market and the potential of the product or idea of the startup. In this thesis, we did not try to control for these variables, for which we would need property data of venture capital firms (such as a database of pitch decks), but instead tried to control for other variables that the

[55] See 2.1.2 Alternative Data in Venture Investments.

literature finds significantly correlated with the success of venture investments (see Appendix 2).

Model Testing

Fitting the newly found mode in the relationship between the performance of venture investments and the *Herding Index (HI)* for the testing sample, we find equivalent results. In fact, we see that even out of training sample we are able to capture some structure about data with a similar R-squared of 0.2074[56].

Final Model

Finally, if we run the regression using all the data avaliable at our disposal[57], we find very solid resutls, with a similar explanatory power (R-squared of 0.189) and an even higher confidence level of 99.75% (see Figure 10a and 10b).

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		10.28
Methond:	Least Squares			Prob (F-stat):		0.0025
N. Observations:	46			Log-Likelihood:		-57.001
R-squared:	0.189			AIC:		118.0
Adj. R-squared:	0.171			BIC:		121.7
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.0108	0.128	0.084	0.933	-0.247	0.268
Herding Index (HI)	0.4288	0.134	3.207	0.003	0.159	0.698
Durbin Watson:	1.985					
Jarque-Brera (JB):	2.79					
Prob (JB):	0.248					

Figure 10a: *Herding Index (HI)* and Success of Venture Investments.

[56] R-squared: 0.2074; Mean Squared Error (MSE): 0.9452; Root Mean Squared Error (RMSE): 0.9722; Mean Absolute Error (MAE): 0.7744.

[57] Arguably, a more meaningful thing to do. Training and testing are mostly used in the presence of ML algorithms and big data that needs to be modelled. However, it can be useful in linear regressions as well, to further solidify the reliability of the results.

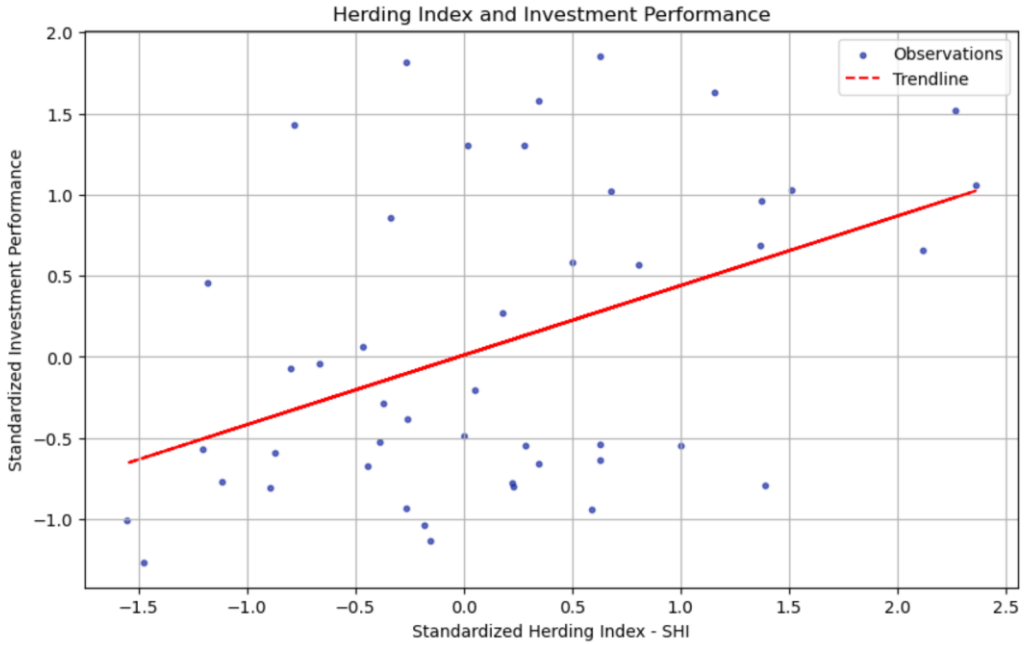


Figure 10b: Graphical Representation of the Model.

Herding and Technological Advancements

We then investigate the relationship between market results and sentiment in different *trend* types, namely industry and technology, to test our second hypothesis.

To do so, we did two separate OLS regressions, to better understand if the relationship between herding and success of venture investments is valid for only technological breakthroughs or not. Regression results are shown by figure 11a and 11b.

OLS Regression Results

Dependant Variable:	Standardized Valuations	F-statistic:	5.986
Methond:	Least Squares	Prob (F-stat):	0.0264
N. Observations:	18	Log-Likelihood:	-21.882
R-squared:	0.272	AIC:	47.76
Adj. R-squared:	0.227	BIC:	49.54

	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	-0.1621	0.206	-0.786	0.444	-0.600	0.275
Herding Index (HI)	0.4091	0.167	2.447	0.026	0.055	0.764

Durbin Watson:	1.809
Jarque-Brera (JB):	1.841
Prob (JB):	0.398

Figure 11a: Technology *Herding Index (HI)* and Success of Venture Investments.

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		6.158
Methond:	Least Squares			Prob (F-stat):		0.0196
N. Observations:	29			Log-Likelihood:		-36.136
R-squared:	0.186			AIC:		76.27
Adj. R-squared:	0.156			BIC:		79.01
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.1246	0.164	0.758	0.455	-0.213	0.462
Herding Index (HI)	0.4133	0.167	2.482	0.02	0.072	0.755
Durbin Watson:	2.464					
Jarque-Brera (JB):	1.366					
Prob (JB):	0.505					

Figure 11b: Industry *Herding Index (HI)* and Success of Venture Investments.

The result indicate that the two models seem to be similar in terms of significance (both around a 95% confidence level, although the technology regression is actually significant with a 90% confidence level).

However, the technology model shows a higher explanatory power (higher R-squared), which testifies its greater strength as a predictor.

Herding and Market Movements

When building the *Herding Index (HI)*, we defined it only in the periods in which the market returns were either exceedingly high or low (booms and busts). These were proxied by the observations of the *Combined Sentiment Score (CSS)*, as many papers found a strong link between this and the short-term movements of the market (which, in the venture capital one, are unobservable).

This assumption is correct, and by investigating the relationship between the volatility of the *Combined Sentiment Score (CSS)* and the success of venture investments when the former is contained between its first and third quartile, we find a confirmation of this, as there seems to be no relationship at all between these two variables in this case.

OLS Regression Results

Dependant Variable:	Standardized Valuations			F-statistic:		0.5331
Methond:	Least Squares			Prob (F-stat):		0.469
N. Observations:	46			Log-Likelihood:		-60.933
R-squared:	0.012			AIC:		125.9
Adj. R-squared:	-0.010			BIC:		129.5
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.2012	0.137	1.466	0.150	-0.075	0.478
Stdev(CSS)	-0.1276	0.175	-0.730	0.469	-0.480	0.225
Durbin Watson:	2.047					
Jarque-Brera (JB):	2.239					
Prob (JB):	0.326					

Figure 12a: Volatility of the *Combined Sentiment Score (CSS)* and Success of Venture Investments in periods of "quiet" markets.

Nonetheless, we find, as anticipated in Appendix 2, that the volatility of the *Combined Sentiment Score (CSS)* seems to be correlated (with a 95% confidence level), in this range, with the survival rate of the startups founded in those years. The regression also presents a strong alpha, as this is the average yearly survival rate, and the volatility of the *Combined Sentiment Score (CSS)* seems to explain a relatively small part (R-squared of 0.1210) of the fluctuations around this value.

OLS Regression Results

Dependant Variable:	Avg. Survival Rate			F-statistic:		6.059
Methond:	Least Squares			Prob (F-stat):		0.0178
N. Observations:	46			Log-Likelihood:		40.791
R-squared:	0.1210			AIC:		-77.58
Adj. R-squared:	0.1010			BIC:		-73.92
	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	0.3956	0.015	26.318	0.000	0.365	0.426
Stdev(CSS)	0.0471	0.019	2.462	0.018	0.009	0.086
Durbin Watson:	2.538					
Jarque-Brera (JB):	2.098					
Prob (JB):	0.350					

Figure 12b: Volatility of the *Combined Sentiment Score (CSS)* and Survival Rates of Startups.

3.5.3. Interpretation of the Results

In general, we can state that our model is significant, thus we can reject the first null hypothesis (HP0[1]) and state that there exists a negative relationship between the performance of venture investments and the amount of herding in investor's opinions.

However, it is important to note how most of the variability in the returns cannot be captured by this behavioural phenomenon (low R-squared), and it is still bound to other factors, most likely those explored in the previous sections[58].

This is coherent with findings in adjacent literature (Wang & Wang, 2018; Ren and Wu, 2018; Zhang et al., 2014; Zhang et al. 2021) who also find this relationship.

For HP0[2], we do not find sufficient results to confirm this. Although it is true that the technology model is able to better capture the performance of venture investments (higher R-squared), the two models are not significant enough (95% and 90% confidence level) to reject the null hypothesis.

A Possible Explanation

An explanation of the above results is that the venture capital industry tends to act in two different ways depending on the market conditions.

In periods of booms and busts, when there is either a lot of hype or disillusion about a *trend*, then the herding phenomenon can significantly impact the quality of the investments made. As investment returns in these times tend to be more dispersed, with both higher highs and lower lows (Chris Yung, 2017), if investors have a wider range of opinions and make bold bets, then there is a higher chance that the extremely small number of potential unicorns will receive funding and emerge.

On the other hand, in periods of relative "quiet," when there is no clear *trend* that is revolutionizing an industry or the whole world, the herding phenomenon is less present, and it does not significantly impact the quality of investments. Here, investors might not try to make the *home runs* but instead focus more on selecting carefully their investments and having a significant return in their portfolio as a whole (Gompers & Lerner, 2003). In this setting, herding might influence (by a relatively small amount) the survival rates of the startups that are born. In fact, at an industry level, if there is a wide variety of investment theses, then more startups will find an investor willing to support them, and therefore have a higher chance of survival.

[58] See 2.1.2 Alternative Data in Venture Investments.

If these explanation holds, then something that could benefit investors and the whole industry, is having a contrarian investment thesis, and stick to it, as advocated by important investors (Thiel, 2014), especially in times of strong market movements.

Conclusion

This thesis investigated the relationship between social media sentiment, herding and the quality of venture investments, an area that has never been explored, up until now. It builds on top of existing literature to create a theoretical model that can identify herding in venture investors' judgements through an innovative *Herding Index (HI)*.

In exploring this new area of research, it also opens the doors for a new source of alternative data, through the development of the YouTube sentiment scraper[59]. Moreover, it re-defines the idea of success in venture investing, by developing a new variable that is better able to capture this concept.

The results found show that there exists a significant negative correlation between this herding phenomenon and the quality of the investments made. However, it does not find enough evidence that this phenomenon is more pronounced for technological breakthroughs, as commonly though.

Moreover, the thesis is not short of fallacies, in which, I hope, future research could find its roots. First, the model can be improved in many ways. As hinted before, the relationship binding investor sentiment with the quality of venture investments is often misaligned in terms of timings. It would be interesting to further investigate this phenomenon, to better understand its causes and consequently improve the *Herding Index (HI)*. Some hints are given during the thesis, but this topic has remained still unexplored.

Furthermore, as Renault points out (2017) it could be interesting to develop a lexicon (dictionary) of VC-specific terms and expressions, with the aim of improving the sentiment analysis and thus the calculation of the *Herding Index (HI)*.

Finally, the model is not taking into consideration the distinction between rational and irrational herding behaviour (Devenow and Welch, 1996), whereby only the second one should imply a positive correlation between the *Herding Index (HI)* and the investment's

[59] See Appendix 3.

quality. This could be impacting the results, and having the ability to discern between the two herding types could significantly improve the *Herding Index (HI).*

Notwithstanding these fallacies, this model has the ambition to become part of a broader hybrid intelligence framework, particularly useful for less experienced investors, which can be used to de-bias their own judgement (thus, the title of this thesis) because, to put it into Kahneman's words:

> *"The way to block biases is simple in principle: recognize the signs that you are in a cognitive minefield, slow down, and think rationally"*
> — *Thinking, Fast and Slow* (2011), p. 417

References

Aerts, K., Matthyssens, P., Vandenbempt, K. (2007). "Critical role and screening practices of European business incubators." *Technovation*, 27: 254-267.

Afful-Dadzie, E., Afful-Dadzie, A. (2016). "A decision-making model for selecting start-up businesses in a government venture capital scheme." *Management Decision*, Vol. 54 (3): 713-734.

Akerlof, G.A., Shiller, R.J. (2009). "Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism." *Princeton University Press*.

Audrino, F., Sigrist, F., Ballinari, D. (2020). "The impact of sentiment and attention measures on stock market volatility." *International Journal of Forecasting*, 36: 334-357.

Bai, S., Yijun, Z. (2021). "Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches." *Systems*. 9(3): 55.

Baker, M., Wurgler, J. (2007). "Investor sentiment in the stock market." *Journal of Economic Perspectives*, 21(2), 129-151.

Banerjee, A. V. (1992). "A simple model of herd behaviour." *The Quarterly Journal of Economics*, 107(3), 797-817.

Barber, B. M., Odean, T. (2001). "Boys will be boys: Gender, overconfidence, and common stock investment." *The Quarterly Journal of Economics*, 116(1), 261-292.

Baum, J., Silverman, B. (2004). "Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups." *Journal of Business Venturing*, 19: 411-436.

Bekiros, S. et al. (2017). "Herding behavior, market sentiment and volatility: Will the bubble resume?" *The North American Journal of Economics and Finance*, 42: 107-131.

Ben Amor, S., Kooli, M. (2024). "Does overconfidence affect venture capital firms' investment?" *Journal of Behavioural and Experimental Finance*, 41.

Blohm, I., et al. (2020). “It’s a Peoples Game, Isn’t It?! A Comparison Between the Investment Returns of Business Angels and Machine Learning Algorithms.” *Journal of Business Venturing*, 35(6), 105969.

Bock, A. J., George, G. (2018). “The role of the founding team in entrepreneurial exits.” *Journal of Business Venturing*, 33(1), 23-37.

Bollen, J., Mao, H., Zeng, X. (2011). “Twitter mood predicts the stock market.” *Journal of Computational Science*, 2(1), 1-8.

Bonein, H., Carat, G. (2020). “The rise of alternative data in venture capital: A new approach to investment.” *Journal of Business Research*, 108, 421-429.

Brown, G. W., Iverson, D., Weisbenner, S. (2012). “Investment sentiment and venture capital flows.” *Journal of Financial Economics*, 106(2), 356-372.

Cao, L. et al. (2023). "Sourcing Investment Targets for Venture and Growth Capital Using Multivariate Time Series Transformer." *Papers 2309.16888*, arXiv.org.

Cao, L. et al. (2022). "Using Deep Learning to Find the Next Unicorn: A Practical Synthesis on Optimization Target, Feature Selection, Data Split and Evaluation Strategy." *Workshop of Multimodal AI For Financial Forecasting* (MuFFin).

Caruso, C. et al. (2015). "Algorithmic Venture Capital." *Stanford.edu*. Reports.

CBInsights (2023). “278 of the biggest, costliest startup failures of all time”. https://www.cbinsights.com/research/biggest-startup-failures/. [Last Access 16/09/2024].

Corea, F., Bertinetti, G., Cervellati, M.E. (2021). "Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors." *Machine Learning with Applications*, 5.

Da Rin, M., Hellmann, T., Puri, M. (2013). “A survey of venture capital research.” *Handbook of the Economics of Finance*, 2(1), 573-648.

Da, Z., Engelberg, J., & Gao, P. (2015). “The sum of all FEARS investor sentiment and asset prices.” *Review of Financial Studies*, 28(1), 1-32.

Dasgupta, A., Prat, A., Verardo, M. (2011). “The price impact of institutional herding.” *The Review of Financial Studies*, 24(3), 892-925.

De Clercq, D. et al. (2006). “An entrepreneur's guide to the venture capital galaxy.” *Academy of Management Perspectives*, 20(3), 90-112.

Deias, A., Magrini, A. (2023). "The Impact of Equity Funding Dynamics on Venture Success: An Empirical Analysis Based on Crunchbase Data." *Economies*, 11(1), 19.

Dellermann, D. et al. (2021). “Hybrid Intelligence.” 10.48550/arXiv.2105.00691.

Devenow, A., Welch, I. (1996). “Rational herding in financial economics.” *European Economic Review*, 40(3-5), 603–615.

Dutta, S. et al. (2021). "A Study on Herd Behaviour Using Sentiment Analysis in Online Social Network." *International Conference on Communication, Control and Information Science*.

Eagle Alpha (2024), https://www.eaglealpha.com/what-is-alternative-data/. [Last access 08/09/2024].

Economic Times (2024). “Average time for startup to turn unicorn has lessend to 5.5 years: Report.” https://economictimes.indiatimes.com/tech/startups/average-time-for-startup-to-turn-unicorn-has-lessened-to-5-5-years-orios-vp/articleshow/106905229.cms?from=mdr. [Last access 08/09/2024].

Forbes, D. P., Milliken, F. J. (1999). “Cognition and corporate governance: Understanding boards of directors as strategic decision-making groups.” *Academy of Management Review*, 24(3), 489-505.

Franke, N. et al. (2008). “Venture capitalists’ evaluations of start-up teams: Trade-offs, knock-out criteria, and the impact of VC experience.” *Entrepreneurship Theory and Practice*, 32(3), 459-483.

Gans, J. S., Hsu, D. H., & Stern, S. (2002). “When does start-up innovation spur the gale of creative destruction?” *RAND Journal of Economics*, 33(4), 571-586.

Gompers, P. A. (1995). "Optimal investment, monitoring, and the staging of venture capital." *The Journal of Finance*, 50(5), 1461-1489.

Gompers, P., Lerner, J. (2003). "Short-Term America Revisited? Boom and Bust in the Venture Capital Industry and the Impact on Innovation." *In Adam B. Jaffe, Josh Lerner, & Scott Stern (Eds.), Innovation Policy and the Economy*, Volume 3. MIT Press.

Gompers, P. A., Lerner, J. (2001). "The venture capital revolution." *Journal of Economic Perspectives*, 15(2), 145-168.

Gompers, P. et al. (2008). "Venture capital investment cycles: The impact of public markets." *Journal of Financial Economics*, 87(1), 1-23.

Greenspan, A. (1996). "Remarks by Chairman Alan Greenspan." Federal Reserve Board, December 5. https://www.federalreserve.gov/boarddocs/speeches/1996/19961205.htm [last access 20/08/24].

Gompers, P.A. et al. (2020). "How do venture capitalists make decisions?". *Journal of Financial Economics*, 135: 169-190.

Gruber, M., MacMillan, I. C., Thompson, J. D. (2008). "Look Before You Leap: Market Opportunity Identification in Emerging Technology Firms." *Management Science*, 54(9), 1652-1665.

Gupta, A., Mudgill, P., Wang, Z. (2019). "Alternative data: Applications in equity investing." *Journal of Investment Strategies*, 8(3), 45-60.

Hirshleifer, D., Shumway, T. (2003). "Good day sunshine: Stock returns and the weather." *The Journal of Finance*, 58(3), 1009-1032.

Hochberg, Y. V., Ljungqvist, A., Lu, Y. (2007). "Whom you know matters: Venture capital networks and investment performance." *The Journal of Finance*, 62(1), 251-301.

Hsu, D. H. (2004). "What do entrepreneurs pay for venture capital affiliation?" *The Journal of Finance*, 59(4), 1805-1844.

Huang, L., Pearce, J. L. (2015). “Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions.” *Administrative Science Quarterly, 60*(4), 634–670.

Justina, D., Helyette, G., Carlo, P., Provetti, A. (2022). "A sentiment analysis approach to the prediction of market volatility." *Frontiers in Artificial Intelligence*, 5.

Kahneman, D., Lovallo, D. (1993). “Timid choices and bold forecasts: A cognitive perspective on risk-taking.” *Management Science*, 39(1), 17-31.

Kaplan, S.N., Sensoy, B.A., Stromberg, P. (2009). “Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies.” *The Journal of Finance*, 64 (1): 75-115.

Kirsch, D., Goldfarb, B., Gera, A. (2009). “Form or substance: the role of business plans in venture capital decision making.” *Strategic Management Journal*, 30: 487-515.

Klotz, A. C. et al. (2014). “New venture teams: A review of the literature and roadmap for future research.” *Journal of Management*, 40(1), 226-255.

Kolanovic, M., Krishnamachari, R. (2017). “Big data and AI strategies: Machine learning and alternative data in trading.” *J.P. Morgan Global Market Strategy*.

Li, J., Gong, Y., & Qing, S. (2023). “Market Sentiment Analysis Based on Social Media and Trading Volume for Asset Price Movement Prediction.” *SpringerLink*.

Li, T., et al. (2023). “Understanding the role of social media sentiment in identifying irrational herding behaviour in the stock market.” *International Review of Economics and Finance,* 87(1), 163-179.

Loughran, T., & McDonald, B. (2011). “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.” *The Journal of Finance*, 66(1), 35-65.

Luss, R., d’Aspremont, A. (2015). “Support vector machine classification with indefinite kernels.” *Mathematical Programming*, 142(1-2), 45-73.

Mason, C., Stark, M. (2024). “What do Investors Look for in a Business Plan?” *International Small Business Journal*, Vol 22 (3): 227-248.

Maxwell, A., Jeffrey, S. (2011). “Business Angel Early Stage Decision Making.” *Journal of Business Venturing*, 26(2):212-225.

Nanda, R., Rhodes-Kropf, M. (2013). “Investment cycles and startup innovation.” *Journal of Financial Economics*, 110(2), 403-418.

Nassirtoussi, A. K. et al. (2014). “Text mining for market prediction: A systematic review.” *Expert Systems with Applications*, 41(16), 7653-7670.

Naughton, J. (2023). “How did Sam Bankman-Fried attract investors? Well, Fomo probably helped.” *The Guardian.*

Nicolas, M.L.D. (2022). "Estimating a model of herding behaviour on social networks." *Physica A: Statistical Mechanics and its Applications*, 604.

Oswald, Y. (2021). “Chasing Home Runs.” https://www.yannickoswald.com/post/homeruns-are-what-matter. [Last access 08/09/2024].

Owusu A.M., Koranteng K.P., Dadzie, J., Dadson, B.A. (2013). "Application of Principal Components Analysis on Venture Capital Investment Decision Process in Ghana." *Journal of Economics and Sustainable Development*, 4(20): 104-108.

Pang, B., Lee, L. (2008). “Opinion mining and sentiment analysis.” *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Peakmetrics, (2024). “The impact of reddit sentiment on stock market prices.” https://www.peakmetrics.com/insights/the-impact-of-reddit-sentiments-on-stock-market-prices. [Last access 15/09/2024].

Prosad, J. M., Kapoor, S., Sengupta, J. (2015). “Theory of behavioral finance.” In *Handbook of research on behavioral finance and investment strategies: Decision making in the financial industry,* IGI Global pp. 24-48.

Qingquan, T., Beibei, L., Xie, D. (2020). “Alternative Data and Artificial Intelligence Techniques: Applications in Investment and Risk Management." *Palgrave Studies in Risk and Insurance*.

Rampell, A. (2015) “Distribution vs. Innovation.” https://a16z.com/distribution-vs-innovation/. [Last access 10/09/2024].

Ren, R., Wu, D. (2018). “An Innovative Sentiment Analysis to Measure Herd Behavior.” *IEEE Transactions on Systems Man and Cybernetics Systems,* 99(1-11).

Renault, T. (2017). “Intraday online investor sentiment and return patterns in the US stock market.” *Journal of Banking & Finance*, 84, 25-40.

Rizk, Y., Mabrouk, M., Otrok, H. (2019). “A novel multi-criteria system for online financial opinion mining based on credibility analysis.” *Future Generation Computer Systems*, 95, 406-418.

Ross, G., Das, S., Sciro, D., Raza, H. (2021). "CapitalVX: A machine learning model for startup selection and exit prediction". *The Journal of Finance and Data Science*, 7: 94-114.

Sahut, J. M., Boulerne, S., Teulon, F. (2019). “The role of alternative data in venture capital investment decisions.” *Technological Forecasting and Social Change*, 142, 87-101.

Saunders, C. (2020). “The Importance of Founder-Market Fit & How To Highlight It While Fundraising.” https://www.forbes.com/sites/jilliancanning/2020/01/15/the-importance-of-founder-market-fit--how-to-highlight-it-while-fundraising/. [Last access 10/09/2024].

Scharfstein, D. S., Stein, J. C. (1990). “Herd behaviour and investment.” *The American Economic Review*, 80(3), 465-479.

Schumaker, R. P., Chen, H. (2009). “Textual analysis of stock market prediction using breaking financial news: The AZFin text system.” *ACM Transactions on Information Systems (TOIS),* 27(2), 1-19.

Schwienbacher, A. (2013). “The entrepreneur’s investor choice: The impact on later-stage firm development.” *Journal of Business Venturing*, 28(1), 72-88

Sewell, M. (2011). “History of the Efficient Market Hypothesis.” *Research Note RN/11/04,* University College London, Department of Computer Science.

Shiller, R. J. (2000). “Measuring bubble expectations and investor confidence.” *The Journal of Psychology and Financial Markets*, 1(1), 49-60.

Siemer, M. (2010). "Behavior under uncertainty and the economics of information." *Journal of Economic Literature*, 48(3): 481-513.

Siering, M., Deokar, A. V., Janze, C. (2017). “The financial sentiment analysis and the impact of alternative data on equity investments: A review.” *Journal of Information Technology*, 32(4), 367-388.

Smales, L. A. (2014). “News sentiment and the investor fear gauge.” *Finance Research Letters*, 11(2), 122-130.

Sourcescrub, (2023). “Research Tools for PE and VC firms.” https://www.sourcescrub.com/uk/post/research-tools-pe-vc-firms. [Last access 16/09/2024].

Sudek, R. (2006). “Angel Investment Criteria.” *Journal of Small Business Strategy*, 17 (2): 89.

Sun, Y. et al. (2022). “A Survey on Alternative Data in Finance and Business: Emerging Applications and Theory Analysis." *Available at SSRN*: https://ssrn.com/abstract=4148628 or http://dx.doi.org/10.2139/ssrn.4148628.

TechCrunch (2024). “Social networks are getting stingy with their data, leaving third-party developers in the lurch.” https://techcrunch.com/2024/02/09/social-network-api-apps-twitter-reddit-threads-mastodon-bluesky/. [Last access 14/09/2024].

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.

The Muse, (2020). “51 Tech and Business Resources to use if you want to be an expert in your field.” https://www.themuse.com/advice/51-tech-and-business-resources-to-use-if-you-want-to-be-an-expert-in-your-field. [Last access 15/09/2024].

Thiel, P. (2014). *Zero to one*. New York: Crown Business.

Tumasjan, A., Braun, R., Stolz, B. (2021). "Twitter sentiment as a weak signal in venture capital financing." *Journal of Business Venturing*, 36(2).

Vanacker, T. et al. (2020). “Signal Strength, Media Attention, and Resource Mobilization: Evidence from New Private Equity Firms.” *Academy of Management Journal*, 63(4).

Wang, G., Wang, Y. (2018). “Herding, social network and volatility.” *Economic Modelling*, Elsevier, vol. 68(C), pages 74-81.

Yin, B., Luo, J. (2018). “How Do Accelerators Select Startups? Shifting Decision Criteria Across Stages.” *IEEE Transactions on engineering management*, Vol. 65 (4).

Yung, C. (2016). “Venture Capital and the Market for Talent during Booms and Busts.” *Review of Finance*, 21(5), 1875-1899.

Zacharakis, A., Shepherd, D. A. (2001). “The nature of information and overconfidence on venture capitalists' decision making.” *Journal of Business Venturing*, 16(4), 311-332.

Zava, M. (2023). “Midas: Get the Golden Investors.” Available at SSRN: https://ssrn.com/abstract=4713531 or http://dx.doi.org/10.2139/ssrn.4713531 .

Zava, M., Caselli, S. (2023). “Venture Capital Networks.” *The Palgrave Encyclopedia of Private Equity*. Available at SSRN: https://ssrn.com/abstract=4433596.

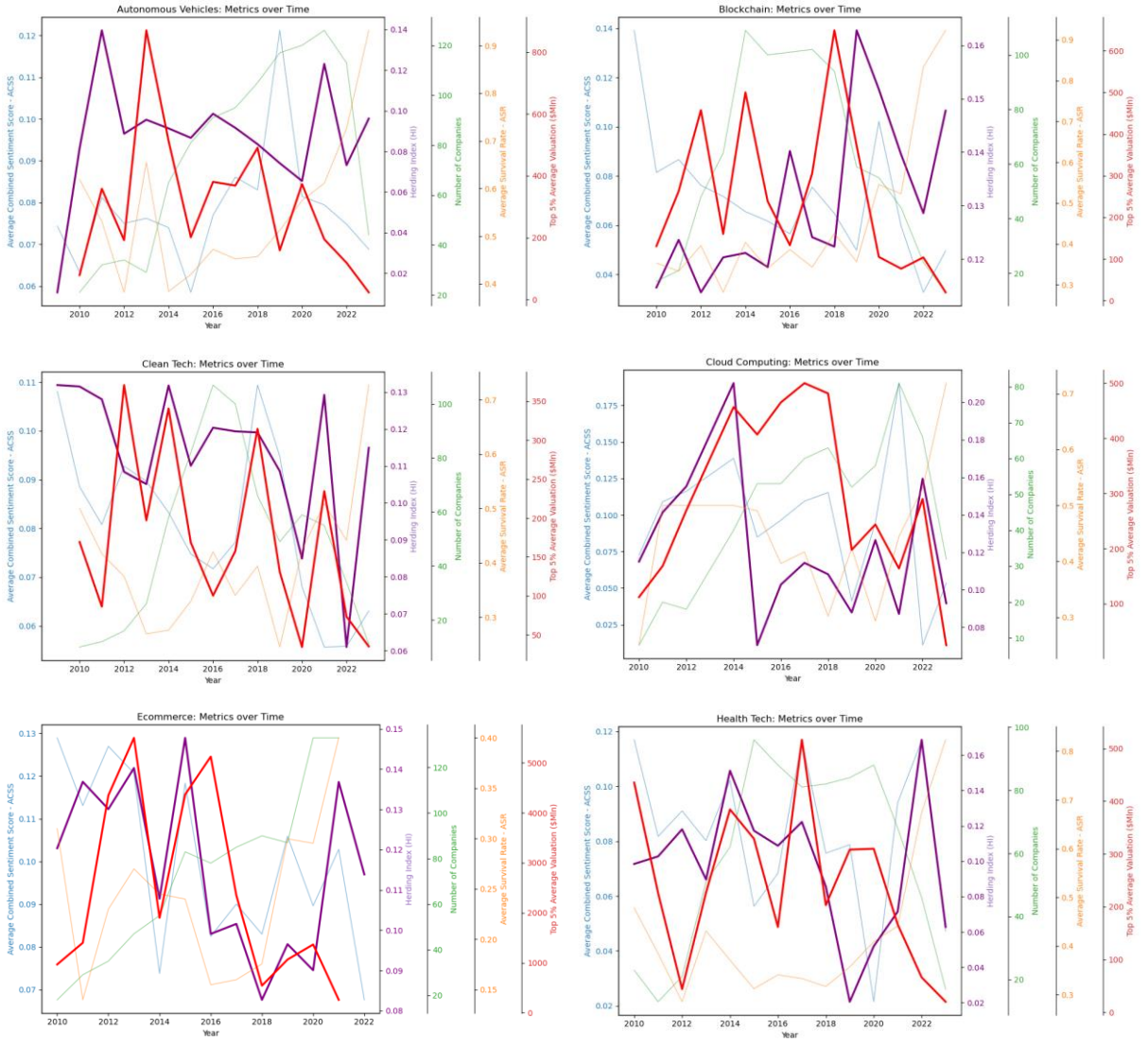
Zhang, J., Jiang, F., Tong, G. (2014). "Venture Capitalists' Herding and the Performance of New Ventures." Management Science, 60(2): 522-537.

Zhang, Q. T., Liu, G. P. (2020). *Alternative Data and Artificial Intelligence Techniques for Trading and Investment*. Wiley.

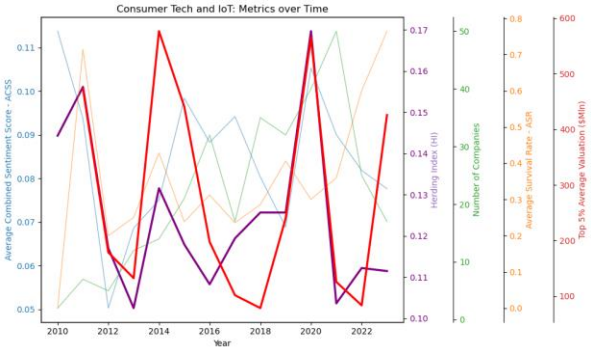
Zhang, R. et al. (2021). "Herd Behaviour in Venture Capital Market: Evidence from China." *Mathematics*, 9.

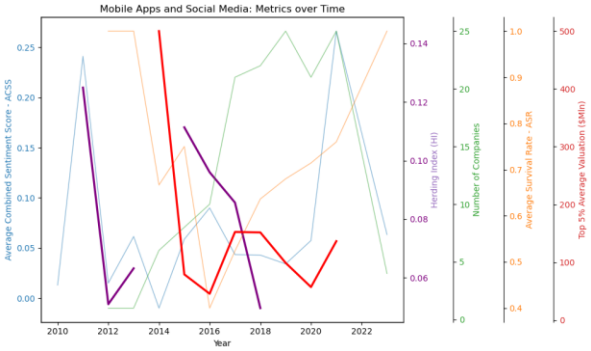
Appendix 1

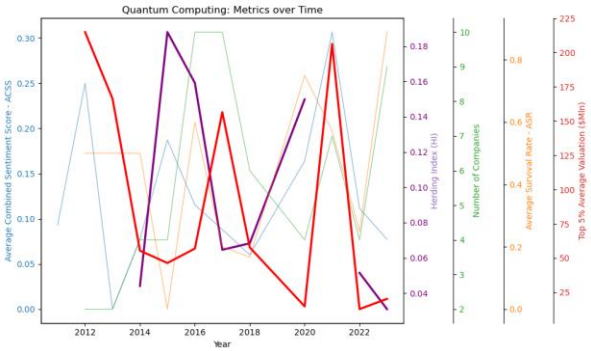
Here are reported the plots that, for each trend, show the average of the top 5% most valued ($Mln) startups founded each year, our dependant variable, and the *Herding Index (HI)*, our independent variable. Together with these two metrics, we also plot (in the background) the *Average Combined Sentiment Score (ACSS)* and the Survival Rate (SR-%)[60].

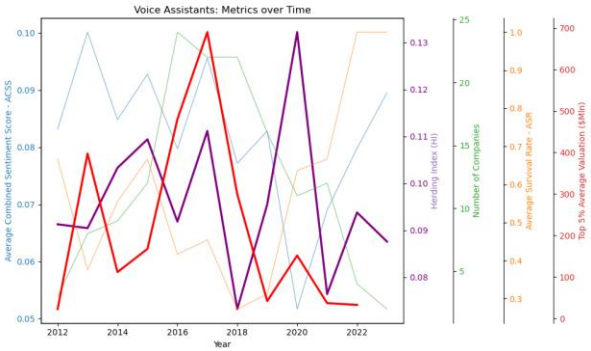


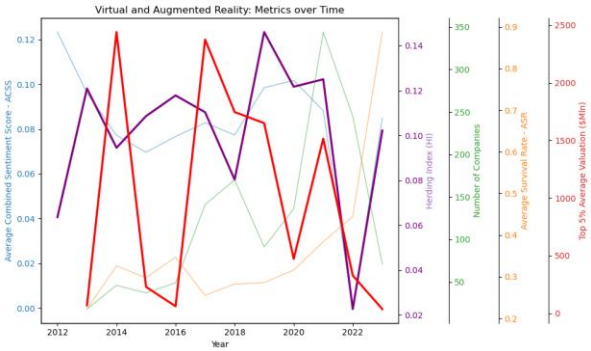
[60] The survival rate is calculated as the % of companies that are still active to date. It is interesting to note how there clearly is a strong positive correlation with this and the *Herding Index (HI)*, hinting to the fact that when there is less herding, safer investments are made (and vice versa). However, this is in not a good "success" metric for VCs and is therefore not the focus of this thesis. Still, it could be of use for future research.


Consumer Tech and IoT: Metrics over Time
Average Combined Sentiment Score - ACSS
Herding Index (HI)
Number of Companies
Average Survival Rate - ASR
Top 5% Average Valuation ($Mln)
Year


Mobile Apps and Social Media: Metrics over Time
Average Combined Sentiment Score - ACSS
Herding Index (HI)
Number of Companies
Average Survival Rate - ASR
Top 5% Average Valuation ($Mln)
Year


Quantum Computing: Metrics over Time
Average Combined Sentiment Score - ACSS
Herding Index (HI)
Number of Companies
Average Survival Rate - ASR
Top 5% Average Valuation ($Mln)
Year


Voice Assistants: Metrics over Time
Average Combined Sentiment Score - ACSS
Herding Index (HI)
Number of Companies
Average Survival Rate - ASR
Top 5% Average Valuation ($Mln)
Year


Virtual and Augmented Reality: Metrics over Time
Average Combined Sentiment Score - ACSS
Herding Index (HI)
Number of Companies
Average Survival Rate - ASR
Top 5% Average Valuation ($Mln)
Year

Appendix 2

Trying to control for IPO market performance or capital inflows in the VC market as literature (Tumasjan et al., 2021) points out, I found no significant relationship, and therefore excluded these kind of control variables from the analysis.

OLS Regression Results

Dep. Variable:	valuation_standardized	R-squared:	0.204
Model:	OLS	Adj. R-squared:	0.141
Method:	Least Squares	F-statistic:	3.252
Date:	dom, 08 set 2024	Prob (F-statistic):	0.0321
Time:	22:16:07	Log-Likelihood:	-52.399
No. Observations:	42	AIC:	112.8
Df Residuals:	38	BIC:	119.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0739	0.144	0.512	0.611	-0.218	0.366
Total_IPOs	0.0601	0.257	0.234	0.816	-0.461	0.581
Total_VC_Inflows_Billion	-0.2103	0.265	-0.792	0.433	-0.748	0.327
std_observations_standardized	0.4216	0.142	2.972	0.005	0.134	0.709

Omnibus:	2.616	Durbin-Watson:	2.047
Prob(Omnibus):	0.270	Jarque-Bera (JB):	2.173
Skew:	0.431	Prob(JB):	0.337
Kurtosis:	2.294	Cond. No.	2.58

Appendix 3: YouTube Sentiment Scraper

Setting up the algorithm

We need to setup the environment in which our algorithm is going to operate.

Import Relevant Libraries

First of all, we import the libraries that we are going to use:

- re: a Python library for working with regular expressions, providing tools for searching, matching, and manipulating text patterns.
- numpy: provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
- pytube: a Python library for downloading YouTube videos and extracting streams from them.
- textblob: a library for processing textual data, offering tools for text analysis like part-of-speech tagging, sentiment analysis, and translation.
- nltk.corpus: a module in the Natural Language Toolkit (NLTK) that provides access to a wide range of corpora and lexical resources useful for language processing tasks.
- youtubesearchpython: a Python library that allows programmatic searching of YouTube videos, channels, and playlists.
- youtube_transcript_api: a Python tool for extracting or generating transcripts (captions) from YouTube videos using AI.

In [1]:
```
import re
import numpy as np
from pytube import extract
from textblob import TextBlob
from youtubesearchpython import *
from youtube_transcript_api import YouTubeTranscriptApi
from datetime import datetime
```

Define Relevant Functions

Now, let's define the functions that we are going to use:

- time_string_to_days(video_bio): allows us to calculate the age of a youtube video (how long ago the video was uploaded)
- clean(transcript_list): allows us to clean the transcript of a youtube video from all irrelevant elements
- video_id(link): allows us to extract the ID of a YouTube video from its link
- join_sentences_with_keyword(raw_list; kw_list): this allows us to extract, for each video transcript, only the relevant sentences (those containing our keywords).

In [2]:
```
def time_string_to_days(video_bio):
    # Define regular expression patterns to extract days, months,
    # and years
    # These is how information in stored in youtube_transcript_api
    seconds_pattern = r'(\d+)\s*second'
    minutes_pattern = r'(\d+)\s*minute'
    hours_pattern = r'(\d+)\s*hour'
    days_pattern = r'(\d+)\s*day'
    months_pattern = r'(\d+)\s*month'
    years_pattern = r'(\d+)\s*year'
    # Extract this information from the string given to us by
    # yotube_transcript_api
    seconds = sum(int(match.group(1)) for match in
                  re.finditer(seconds_pattern, video_bio))
    minutes = sum(int(match.group(1)) for match in
                  re.finditer(minutes_pattern, video_bio))
    hours = sum(int(match.group(1)) for match in
                re.finditer(hours_pattern, video_bio))
    days = sum(int(match.group(1)) for match in
               re.finditer(days_pattern, video_bio))
    months = sum(int(match.group(1)) for match in
                 re.finditer(months_pattern, video_bio))
    years = sum(int(match.group(1)) for match in
                re.finditer(years_pattern, video_bio))
    # Calculate the total number of days
    total_days = (seconds / 84600 + minutes / 1440
                  + hours / 24 + days + months * 30
                  + years * 365)
    return total_days

def clean(transcript_list):
    # Regular expression to match words inside square brackets,
    # eg [APPLAUSE]
    bracket_pattern = r'\[.*?\]'
    # Regular expression to match dots and commas.
    dot_pattern = r'\.'
    comma_pattern = r','
    # Let's now clean the transcript
    cleaned_transcript_list = []
    # Iterate through each dictionary in the list
    for item in transcript_list:
```

```
        # Check if the 'text' value contains words inside
        # square brackets
        if re.search(bracket_pattern, item['text']):
            # If it does, do not include this dictionary
            # in the cleaned list
            continue
        else:
            # Otherwise, remove dots and commas from the
            # 'text' value
            cleaned_text = re.sub(dot_pattern, '',
                                  item['text'])
            cleaned_text = re.sub(comma_pattern, '',
                                  cleaned_text)
            item['text'] = cleaned_text
            cleaned_transcript_list.append(item)
    return cleaned_transcript_list

def video_id(link):
    query = urlparse.urlparse(link)
    # There are many ways in which the link can be built,
    # we need to decompose each of them differently to
    # extract the video ID
    if query.hostname == 'youtu.be':
        return query.path[1:]
    if query.hostname in ('www.youtube.com', 'youtube.com'):
        if query.path == '/watch':
            p = urlparse.parse_qs(query.query)
            return p['v'][0]
        if query.path[:7] == '/embed/':
            return query.path.split('/')[2]
        if query.path[:3] == '/v/':
            return query.path.split('/')[2]
    return None

def join_sentences_with_keyword(raw_list, kw_list):
    # For each sentence in the raw_list, we look if there
    # is a keyword (contained in kw_list). If there is,
    # we store that sentence, the previous and the next
    # one (for context) in a new list (result_list) so
    # that we can analyse it later. If the sentence that contains
    # the keyword happens to be after a sentence that already
    # contained a keyword (check_list), the two shall be joined
    # together as most likely they are part of the same argument.
    result_list = []
    check_list = []
    i = 0
    n = len(raw_list)
    while i < n-1:
        if raw_list[i] in check_list:
            s = len(result_list)-1
            result_list[s] = result_list[s] + raw_list[i+1]
            i = i+1
        else:
```

```
            if i == 0:
                if any(keyword in raw_list[i]
                       for keyword in kw_list):
                    joined_sentences = (raw_list[i]
                                        + raw_list[i+1])
                    result_list.append(joined_sentences)
                    check_list.append(raw_list[i+1])
                    i = i+1
                else:
                    i = i+1
            elif n-1 > i > 0:
                if any(keyword in raw_list[i]
                       for keyword in kw_list):
                    joined_sentences = (raw_list[i-1]
                                        + raw_list[i]
                                        + raw_list[i + 1])
                    result_list.append(joined_sentences)
                    check_list.append(raw_list[i+1])
                    i = i+1
                else:
                    i = i+1
            else:
                break
    return result_list
```

Define the Research Interest

Choice of the Topic

The *topic of interest* is defined through a list (kw_list) containing the keywords that are likely to be found when the topic of interest in the subject of the argument. It is essential to define these words in all the possible formats in which they might appear, because our algorithm will look for an exact match of these in the youtube transcript, as seen above in the join_sentences_with_keywords() function.

In [3]:
```
# In this case we want to investigate Virtual and Augmented Reality
kw_list = ['virtual reality', 'Virtual Reality', 'Virtual reality',
           'VR', 'virtual-reality', 'Virtual-Reality',
           'Virtual-reality', 'VR', 'augmented reality',
           'Augmented Reality', 'Augmented reality', 'AR',
           'augmented-reality', 'Augmented-Reality',
           'Augmented-reality', 'mixed reality', 'Mixed reality',
           'Mixed Reality', 'mixed-reality', 'Mixed-reality',
           'Mixed-Reality']
```

Choice of the time period

The *time period of interest* is then inserted manually to be then converted into appropriate format and stored in the variables "start_period_date" and "end_period_date"

In [ ]:
```
start_date_years = int(2010)
end_date_years = int(2024)
today = datetime.now()
start_period_date = (today - datetime(start_date_years,
                                      1, 1)).days
end_period_date = (datetime(end_date_years,
                            12, 31) - today).days
```

Choice of the Channel

The *channel of interest* is defined through the channel ID (channel_id_1), from which we extract all the videos ever uploaded using the function Playlist(playlist_from_channel_id())

In [5]:
```
channel_id_1 = "UCkkhmBWfS7pILYIk0izkc3A"
playlist_1 = Playlist(playlist_from_channel_id
                      (channel_id_1))
```

Build the Algorithm

We are now going to use all the libraries, functions and inputs collected to create an algorithm that will analyse the YouTube sentiment. It will be made out of two components: the *Video Scraper* and the *Transcript Analyser*

Video Scraper

This piece of code retrieves all the youtube videos from the channel through the function .getNextVideos() in chronological order (from the most recent to the oldest) until it reaches the desired historical starting point of the analysis. All the information about the videos is stored in playlist_1.

In [ ]:
```
start_date_tracker1 = 0
while playlist_1.hasMoreVideos:
    # We will download vidoes until the time of the age of the
    # last downloaded video (start_date_tracker_1) is less than
    # the age of interest (start_period_date)
    if start_date_tracker1 < start_period_date:
        try:
            playlist_1.getNextVideos()
            playlist_1_videos = playlist_1.videos
            s = len(playlist_1_videos)-1
```

```
            video_s = playlist_1_videos[s]
            intermediate1_s = video_s.get('accessibility', None)
            Intermediate2_s = intermediate1_s.get('title', None)
            # The time is embedded in the metadata of the video,
            # thus we need to extract it
            pattern = r'views(.*)$'
            match = re.search(pattern, Intermediate2_s)
            if match:
                period_str = match.group(1).strip()
                # Convert the period str into a total number of days
                total_days = time_string_to_days(period_str)
                start_date_tracker1 = total_days
                print(f'Videos Retrieved for Channel 1: {len(
                    playlist_1.videos)}')
        except Exception:
            print('Error in retreiving videos. Process continues.')
            continue
    else:
        break
```

We then extract from these videos some relevant information:

- The lenght of each video (times1) and its link (links1)
- The date in which the video was uploaded (periods_1) which we need in order to then combine this data with the market data of venture invetment

In [ ]:
```
i = len(playlist_1.videos)-1
times1 = []
links1 = []
periods1 = []
while i >= 0:
    video_i = playlist_1.videos[i]
    # Durations and links are easily available in the playlist
    times1.append(video_i.get('duration', None))
    links1.append(video_i.get('link', None))
    intermediate1_i = video_i.get('accessibility', None)
    Intermediate2_i = intermediate1_i.get('title', None)
    # The date of the video, just like before, needs to
    # be extracted from a string
    pattern = r'views(.*)$'
    match = re.search(pattern, Intermediate2_i)
    if match:
        # If the video ha a duration, then we consider it...
        period_str = match.group(1).strip()
        total_days = time_string_to_days(period_str)
        periods1.append(total_days)
    else:
        # ...otherwise we will discard it
        links1 = links1[:-1]
        times1 = times1[:-1]
    i = i-1
```

```
print('The total duration of the videos analysed for this channel is',
       sum(times1))
```

Transcript Analyser

Once we have retrieved all the relevant videos, we need to download, clean, and then analyse the transcripts.

As a preliminary step, convert the links into video IDs, using the previously defined function extract.video_id(), so that they can be downloaded using the imported libraries.

In [ ]:
```
IDs_1 = []
for url in links1:
    IDs_1.append(extract.video_id(url))
```

At this point we are ready for the transcript download and analysis.
First, we download it using the just created list IDs_1 and the imported libraries.

Then we need to clean the transcript. This is achieved thorugh a series of steps as follows:

a. Use clean()
We use the previously defined clean() function to delete all the elements that we don't want to be part of the analysis

b. Punctuation
We remove and add back all punctuation. Unfortunately, the great majority of YouTube videos does not have correct punctuation, either because it is entirely missing or because the auto-generated subtitles do not seem to manage well this task. Therefore, we need to re-create punctuation from scratch in order to distinguish between sentences in the text, so that we can later analyse them.

To do so we first need to understand how subtitles work. Each subtitle is a line that contains the text displayed in the video and some other metadata. More specifically, it contains, for each line of the subtitle, the time in which it appears (start) and stays on the screen (duration).
The time in which the line stops to be shown, can be derived as:

> $$end(i) = start(i) + duration(i)$$

But this would be incorrect, as this duration is not the actual time in which the transcript line is spoken, but the time in which it will be visible on the screen. This

means that it is not accurate, as for example the subtitle may remain on the screen more than necessary if another one is instantly shown (because the first one is shifted upwards) or because there is a long pause afterwards (and it does not disappear). Thus, i chose to calculate the ending time as:

> $end(i) = start(i+1)$ - and -
> $duration(i) = start(i+1) - start(i)$

In this way, we are able to then understand how much time it takes for each line of the transcript to be spoken.

Once understood this, we can separate senteces:

- Calculated the average "time per letter" of a certain video by dividing the duration over the total number of letters

  > $$\text{duration per letter} = \frac{\text{total duration}}{\text{total number of letters}}$$

- Distinguish sentences based on the time it takes for a subtitle line to be spoken relative to this metric, always considering the number of letters in the line

There are a number of hypothesis I had to make to apply this logic, and these can be points of improvements for further developments: - Phrases do not stop mid-subtitle: this is almost always true (approx 95% of the time) - Every speaker talks with the same pace: this is a strong assumption, but it is also true that it does not influence significantly the results.

c. Sentiment Analysis
Finally, we use textblob to calculate polarity, subjectivity and combined score. We also store information about the date of the observation and about how many observations were made in that date, so that we can view the time evolution of these observations

In [ ]:
```
# We create the lists where we will store the
# results of our analysis
Sent_Combined_Score = []
Sent_Pol_Score = []
Sent_Subj_Score = []
n_obs1 = []

r = len(IDs_1)-1
while r >= 0:
    try:
        # Download, clean and add punctuation to the sentences
        transcript = YouTubeTranscriptApi.get_transcript(IDs_1[r])
```

```
        cleaned_transcript = clean(transcript)
        # Extract the start and the duration of each line in
        # the transcript
        start = [line['start']
                 for line in cleaned_transcript]
        duration = [line['duration']
                    for line in cleaned_transcript]
        # The time in which the line ends is calulated
        # as the sum of these elements
        end = [x + y for x, y in zip(start, duration)]
        i = 0
        n = len(end) - 2
        diff = []
        while i < n:
            # Calculate the times that interlap between lines,
            # called "diff" in the following
            diff.append(start[i + 1] - end[i])
            i = i + 1
        # Collect all the durations in a list as well
        dur = []
        n = len(start)-1
        i = 0
        while i < n:
            dur.append(start[i+1]-start[i])
            i = i+1
        # Since dur has one element less than the transcript
        # (the last line has a starting point, but since there
        # is no line next, it is not possible to determine
        # its ending point), to make them both of the appropriate
        # lenght I shall remove the last element of the
        # transcript (assuming it does not contain any
        # essential information, which is reasonable
        # since very often this is just a final greeting)
        n = len(cleaned_transcript)-1
        del cleaned_transcript[n]
        # Now calculate the duration per character, our
        # key metric for inserting commas and dots in the text.
        # First, we create a list with the length of each world
        words = [line['text'] for line in cleaned_transcript]
        words_length = []
        for word in words:
            words_length.append(len(word))
        # Then we calculate the average duration per
        # character of each word
        dur_per_char = []
        n = len(words_length)
        i = 0
        while i < n:
            dur_per_char.append(dur[i]/words_length[i])
            i = i+1
        # Finally, we calculate keys statistics to place commas
        # and dots in the text
        comma_dur = np.quantile(dur_per_char, 0.65)
```

```
        dot_dur = np.quantile(dur_per_char, 0.90)
        # We then create a loop that adds either commas or dots
        # based on these metrics. In this part of the code,
        # although I left it so that one could add also dots
        # and try to reason on them, I will only use commas
        # for simplicity.
        cleaned_text = [line['text'] for line in cleaned_transcript]
        n = len(cleaned_text)-1
        i=0
        while i < n:
            if diff[i] > 0:
                cleaned_text[i] = str(cleaned_text[i] + ',')
                i = i + 1
            elif dot_dur > dur_per_char[i] > comma_dur:
                cleaned_text[i] = str(cleaned_text[i] + ',')
                i = i + 1
            elif dur_per_char[i] > dot_dur:
                cleaned_text[i] = str(cleaned_text[i] + ',')
                i = i + 1
            else:
                i = i + 1
        # Now, we clean the text and put it in a list format,
        # where each element is a sentence
        cleaned_text = ''.join(cleaned_text)
        raw_list = cleaned_text.split(',')
        raw_list = [sentence.strip()
                    for sentence in raw_list if sentence.strip()]
        # Only the sentences with keywords will be analysed
        filtered_text = join_sentences_with_keyword(raw_list, kw_list)
        # Finally, raw sentiment scores are calculated and saved
        polarity_scores = []
        subjectivity_scores = []
        for sentence in filtered_text:
            analyse_text = TextBlob(sentence)
            polar_score=analyse_text.sentiment.polarity
            subj_score=analyse_text.sentiment.subjectivity
            Sent_Pol_Score.append(polar_score)
            Sent_Subj_Score.append(subj_score)
            Sent_Combined_Score.append(polar_score*subj_score)
        print("Element ", r, " processed correctly")
        n_obs1.append(len(filtered_text))
        r = r-1
    except Exception:
        # If there is an error, we will skip that video
        # and go on to the next one
        del periods1[r]
        print("Error in processing the ", r, "th element")
        r = r-1
```

Finally, store all the results in separate folders, so that they are easily available for further elaborations

In [ ]:
```
with open(r'C:\File_Path', 'a') as writer:
    for score in Sent_Combined_Score:
        writer.write(str(score) + ",")
with open(r'C:\File_Path', 'a') as writer:
    for score in Sent_Pol_Score:
        writer.write(str(score) + ",")
with open(r'C:\File_Path', 'a') as writer:
    for score in Sent_Subj_Score:
        writer.write(str(score) + ",")
with open(r'C:\File_Path', 'a') as writer:
    for period in periods1:
        writer.write(str(period) + ",")
with open(r'C:\File_Path', 'a') as writer:
    for obs in n_obs1:
        writer.write(str(obs) + ",")
```