CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Department of Economic Studies

Master's Thesis

2023

Toghrul Niyzali

Charles University

Faculty of Social Sciences

Institute of Economic Studies

MASTER THESIS

Artificial intelligence in Behavioral Finance

Author: Toghrul Niyazli

Study Programme: Master in Economics and Finance

Supervisor: İlgar İsmayilov M.A.

Year of the defense: 2023

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on Toghrul Niyazli

7/13/2023

References

Toghrul Niyazli: *Artificial intelligence in Behavioral Finance.* Praha, 2023. 73 pages. Master's thesis (Mgr.). Charles University, Faculty of Social Sciences, Institute of Economic Studies. Department of Economics and Finance. Supervisor İlgar İsmayilov M.A.

Length of the thesis: Number of characters with spaces 120860, 60 Pages.

Abstract

Since the dot.com bubble burst in 2001, financial markets have been plagued by extreme volatility caused by investors emotions and irrational behaviors. Since the financial market has become more complex and tech oriented, stock market sentiment has been a key determinant of large market movements. In this context, it is worth to look into the role of emotions such as fear, greed, and excitement in influencing investment decisions. In addition, newly developed artificial intelligence techniques can be utilized to record and collect data from online sources to conclude collective irrational human behaviors. Behavioral finance is a newly developed discipline that examines the impact of psychological elements on decision-making processes when individuals encounter uncertain situations. This thesis aims to analyze the influence of behavioral finance phenomena including Herding, Anchoring, and Overconfidence biases on the investment decision-making practices within the context of the United States stock market. One of the study goals is to investigate whether sentiment of public, can be a rational tool impacting on investors' decision-making process or not. The investigation is prompted by the potential of Artificial Intelligence and Machine Learning to accurately analyze vast amounts of data and make judgments about the sentiment of financial markets. Scraping and crawling newspapers yielded primary data for analysis. To conduct the above stated analysis, we constructed a new variable, the sentiment factors which stand in core of our analysis. The research proposes that the integration of extensive behavioral data sets, such as stock news and financial stock data, alongside respective web search query volumes can offer novel insights into various phases of extensive collective decision-making processes. This amalgamation enables the identification of collective irrational behavior, commonly known as herding in behavioral finance. The overall goal is to provide the guidelines for constructing portfolios. The sentiment factor and traditional style factors were explored to find the impact of public sentiment on S&P 500 stock prices. The results are explained in a simplified manner and are supported by concrete, practical illustrations developed by using Python and R Studio software packages.

Title: Artificial intelligence in Behavioral Finance

Keywords: Artificial Intelligence, Behavioral Finance, Sentiment Analysis, Machine learning, Python.

Author’s e-mail: 52083211@fsv.cuni.cz

Supervisor’s e-mail: 96719575@fsv.cuni.cz

Abstrakt

Od prasknutí dot.com bubliny v roce 2001 jsou finanční trhy zasaženy extrémní volatilitou způsobenou emocemi a iracionálním chováním investorů. Jelikož se finanční trh stal složitějším a více technologicky orientovaným, sentiment na akciovém trhu se stal klíčovým determinantem velkých tržních pohybů. V tomto kontextu je vhodné se zaměřit na roli emocí, jako jsou strach, chamtivost a nadšení, které ovlivňují investiční rozhodnutí. Navíc mohou být využity nově vyvinuté techniky umělé inteligence pro záznam a sběr dat z online zdrojů k zjištění kolektivního iracionálního lidského chování. Behaviorální finance jsou nedávno vyvinutou disciplínou, která zkoumá dopad psychologických prvků na rozhodovací procesy, když se jednotlivci setkávají s nejistými situacemi. Tato diplomová práce se snaží pochopit dopad konceptů behaviorálních financí, jako jsou shlukování, kotvení a přeceňování, na rozhodovací proces investora, konkrétně na americkém trhu, pomocí technik umělé inteligence. Jedním z cílů studie je zjistit, zda může být sentiment veřejnosti racionálním nástrojem ovlivňujícím rozhodovací proces investora. Toto vyšetřování je motivováno potenciálem umělé inteligence a strojového učení přesně analyzovat velké množství dat a vytvářet úsudky o sentimentu finančních trhů. Vytěžování a prohledávání novin poskytlo primární data pro analýzu. K provedení výše uvedené analýzy jsme vytvořili novou proměnnou, sentimentové faktory, které stojí v jádru naší analýzy. Výzkum navrhuje, že integrace rozsáhlých souborů behaviorálních dat, jako jsou akciové zprávy a finanční akciová data, spolu s příslušnými objemy vyhledávacích dotazů na webu, může nabídnout nové pohledy na různé fáze rozsáhlých kolektivních rozhodovacích procesů. Tato kombinace umožňuje identifikaci kolektivního iracionálního chování, které je v behaviorálních financích známo jako shlukování. Celkovým cílem je poskytnout pokyny pro sestavení portfolií. Sentimentový faktor a tradiční styl faktory byly zkoumány, aby se zjistil dopad veřejného sentimentu na ceny akcií S&P 500. Výsledky jsou vysvětleny zjednodušeným způsobem a jsou podpořeny konkrétními, praktickými ilustracemi vyvinutými pomocí softwarových balíčků Python a R Studio.

Title: Umělá Inteligence v Behaviorálních Financích

Klíčová slova: Umělá inteligence, behaviorální finance, analýza sentimentu, strojové učení, Python.

E-mail autora: 52083211@fsv.cuni.cz

E-mail vedoucího: 96719575@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

AI	Artificial Intelligence
API	Application Programming Interfaces
ANN	Artificial neural networks
APT	Arbitrage Pricing Theory
AWS	Amazon Web Service
CAPM	Capital Asset Pricing Model
EMH	Efficient Market Hypothesis
FFCPS	Fama-French-Carhart-Pastor-Stambaugh
ML	Machine learning
MPT	Modern Portfolio Theory
PoS	Parts of Speech
Python	Programming language
S&P500	Standard & Poor's 500 Index
GMV	Global Minimum Variance
GEM3	Barra Global Equity Model
VIF	Variance Inflation Factors
VIX	Volatility Index

Master's Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Author:	**Bc. Toghrul Niyazli**	Supervisor:	İlgar İsmayilov M.A.
E-mail:	52083211@fsv.cuni.cz	E-mail:	96719575@fsv.cuni.cz
Phone:	+420 774175008	Phone:	
Specialization:	MEF	Defense Planned:	2023 September 21.

Proposed Topic:

Artificial intelligence in Behavioral Finance

Motivation:

Behavioral finance[1] studies investor's psychology of financial decision-making. For a considerable duration, scholars and practitioners have adheredto classical finance theories that postulate the necessity for rationality in market operations and investor decision-making. The psychological aspect of investing is critical for investors to make a final decision in the financial market. Within the sphere of behavioral finance, critical elements such as emotional factors, greed, fear, individual perceptions, and cognitive biases are scrutinized due to their notable influence on investment decision-making in financial markets. Investors should employ effective strategies to maintain consistent profitability in the fast-changing competitive financial market. Behavioral finance is a broad subject and I am specifically interested in sentiment analysis[2] in context of behavioral finance to evaluate sentiment of investors and the way it impacts stock market prices**.** The advent of the COVID-19[3], caused lockdowns and usage of social networks increased significantly starting from 2019. Social networks become common destination for many of us, as people's interaction increased in social media, therefore data analytics companies have a better chance and ability to collect information about people. Data is collected through your device, google searches[4], and things you like and post. The data later on brought into service to make a sentiment and textual analysis to assess the firm-level and market-level behavior[5] and performance in the financial market.

[1] See https://www.jstor.org/stable/3216841
[2] See https://www.jstor.org/stable/j.ctv3t5r09.20
[3] See https://www.jstor.org/stable/resrep25198
[4] See https://www.loyola.edu/academics/emerging-media/blog/2017/3-ways-that-social-media-knows-you-better-than-your-friends-and-family-do
[5] See https://slate.com/business/2015/04/bot-makes-2-4-million-reading-the-web-meet-the-guy-it-cost-a-fortune.html

Artificial Intelligence[6] (AI) has been instrumental in the advancement of behavioral science. Proponents of using artificial intelligence (AI) to make financial decisions claim that utilizing AI helps to solve complex problems, initiate strategic changes, evaluate risks in decision-making, bias avoidance, rational character analysis, and assists to make utilitarian productive investments. Artificial intelligence is also known as machine learning because it is concerned with task that will be performed by machines rather than by humans. Sentiment and textual analysis were carried out artificial intelligence techniques by using rich libraries of Python[7] programing language. Purposes of the paper to show the role of artificial intelligence in behavioral finance, specifically in investors' and traders' decision-making process, algorithms used for sentiment analysis measures uncertainty and actual state of the market. A factor-based model that incorporates sentiment data from news sources to analyze an effect of behavioral finance on stock prices of S&P 500 components can provide valuable insights into the stock market. The sentiment data[8] from a large population can help identify its impact on stock prices. To build this model, the data was first gathered and cleaned, sentiment indicators were developed through feature engineering, and the model was constructed using machine learning methods such as decision trees and neural networks. Large companies and funds[9] use Machine Learning for portfolio allocation and rebalancing, utilizing cost effective techniques to advance its position in the competitive market. Large investment funds employed Machine Learning increased fund efficiency and profitability, a survey[10] from 2019 found that 98% of investors use digital sources to investigate and conduct research. Algorithmic investment services accounting for nearly 90% of the financial market. In this context, factor-based models[11] that incorporate sentiment data have been developed to gain insights into market sentiment (Houlihan & Creamer, 2017; Kim & Kim, 2014; Makrehchi, Shah, & Liao, 2013; Pineiro-Chousa, Lopez-Cabarcos, & Perez-Pico, 2016). For instance, Deutsche Börse (DBAG) has started using artificial neural networks and classification models, powered by artificial intelligence, to collect data from digital sources and build asset valuation models (Glantz & Kissell, 2014). These developments are part of a well-documented positive trend in the application of artificial intelligence and behavioral finance to investment.

1.Hypothesis: Artificial intelligence-based sentiment index is a rational tool for investors.

2.Hypothesis: While markets interact with investors, an inherent mapping exists between investor sentiment and market conditions that reveals future market trajectories.

3.Hypothesis: Investment decision is influenced by psychological and emotional factors.

The integration of artificial intelligence (AI) into finance has brought about many changes in various aspects such as investing, credit scoring, regulatory compliance, market research, and customer support. In particular, this paper will focus on the impact of AI on investing activity. One key application of AI in investing is the use of textual sentiment analysis to model market participant sentiment. This approach leverages language processing software

[6] See https://aws.amazon.com/free/machine-learning/
[7] See https://www.r-project.org/
[8] See https://www.refinitiv.com/en
[9] Robo advisors include Betterment, WealthFront, WiseBanyan, https://www.blackrock.com/aladdin/offerings/aladdin-overview
[10] See https://www.brunswickgroup.com/digital-investor-survey-2020-i15237/
[11] See https://medium.com/wright-research/factor-models-744e17e5d0e5

to analyze news and social media data, in order to determine the prevailing sentiment towards a particular asset or market. I have to confess that research is complicated by a lack of data, first because the phenomenon is relatively new, and second as finance companies and hedge funds use the techniques are often privately, and those are not required to report their returns. Textual sentiment analysis will be used to model market participant confidence and Google trends will be utilized to track public interest. The data will be processed in the context of behavioral finance and an integrated sentiment factor will be developed to enhance traditional factor models. Investment factors refer to specific characteristics or traits of a security, such as value, growth, quality, momentum, etc. Traders and investors have been using these strategies for decades, and well-known traders like Jesse Livermore, Warren Buffet, and George Soros have all been associated with different investment factors. Livermore's approach was trend-following, Buffet's was value and quality, and Soros's was momentum. Building **four** traditional style factor portfolios and one alternative factor portfolio using news **sentiment information** from ML Analytics. This is achieved through a factor mimicking portfolio optimization process, which creates a long-short market-neutral portfolio by neutralizing unwanted factor exposures.

Factorizing a sentiment index involves the following steps:

1. Gather news sentiment data for each security from ML Analytics.
2. Use the data to calculate sentiment scores for each security.
3. Create portfolios considering sentiment scores based on each security.
4. Include it univariate and bivariate portfolio analysis.
5. Test the performance of the portfolio over a historical period to evaluate its potential.

This approach seeks to identify the best individual securities based on alternative factor portfolio exposures.

The results should be correlated with major benchmark indices due to the behavior of market participants. This is because "big data" originates from human interactions in the internet over various social media. Market sentiment analysis is used to understand the correlation between major trends in our social world and their impact on the stock market, which serves as a proxy for our sentiment model. For instance, researchers such as (D'Souza, Ribeiro, & da Silva, 2016) have demonstrated that certain finance-related internet page visits and search statistics tend to correlate with market movements and have been used as proxies for market sentiment analysis. Another example is (Preis, Moat, & Stanley, 2013) who used Google Trends data to construct a sentiment index and found that it was able to predict stock market returns in the UK and US. These studies demonstrate the potential usefulness of internet search activity as a proxy for market sentiment and highlight the importance of considering behavioral factors in investment decision-making. The correlation between major trends in our social world and their impact on the stock market has been explored to help to understand the behavior of market participants.

The objective of this is to enhance investment strategies and mitigate the impact of investor bias by incorporating insights from behavioral finance theories. We are trying to bring light of behavioral finance findings to help investors to design more intuitive algorithms to construct their profitable strategy.

Expected Contribution:

I expect that the model developed in this work using the advantages of AI characteristics incorporated, will be able to obtain signals that generate either negative or positive market responses. In the root of my motivation comes why economists began studying real people to assess the validity of rational behavior, Machine Learning in Finance is not only about prediction but also interpretation of data both qualitative and quantitative, to find out which input variables are more crucial in predicting that target variable as I assume the use of AI in the psychological level biasness can be reduced, and accuracy can be increased.

AI it has been phenomenon since late 20th century, its result is not conclusive, as I hope there is more space to do a research. As we know its complexity and structure does not produce the same result each time, simply the model learns and improves, that is why it produces different result each time, for example, in certain finance related web site search in Google search engine (interaction with human) creates large data set, which people cannot detect such patterns without employing AI. As we witnessed many times AI named alongside Big Data, it is the one of the reasons, we will try to incorporate Big Data as market sentiment. Including these variables in ANN model, I believe it will produce more accurate human behavior patterns which further used to predict next move (Markov decision process). I have built **four** traditional style factor portfolios and one alternative factor-based portfolio using news sentiment information from ML Analytics. It is reported to produce an annualized return of over 400 basis points, even when the portfolio is rebalanced on a monthly basis. This suggests that the sentiment index can provide a significant return for investors, without adding additional risk associated with traditional risk factor.

Outline:

Motivation: The investors sentiment can be interpreted in qualitative way and text analysis can identify hidden human patterns.

Studies on AI demand: News text analysis can generate informative signals for investors

Data: I will explain how I will collect data estimating the accuracy and precision.

Methods: I will explain behavioral finance concept and process the input data accordingly.

Results: I will discuss my baseline regressions and test the outcome with GEM3 model. Concluding remarks: I will summarize my findings and their implications for policy and future research.

Problems: Data harvesting (code of conduct and data privacy)

Introduction

Information is a pivotal factor in the decision-making process. However, obtaining a comprehensive and complete set of information can be challenging, leading to circumstances in which individuals and organizations must make decisions based on insufficient or inaccurate information. Despite these challenges, individuals and organizations are still required to make decisions, which may involve weighing different factors, making assumptions, or taking calculated risks. As a result, decision-making processes are subject to heightened levels of uncertainty and risk. May involve the use of heuristic or probabilistic approaches to weigh available information and select a course of action. All the uncertainties surrounding a choice may lead to an unwanted outcome, therefore understanding people and how they behave is essential in finance. Advances in neuroscience are helping to decode our complex brain and have led to the development of sophisticated approaches in behavioral finance [1]. Although it has been helpful for economists to utilize the idea of rationality [2,3] as the foundation of their theories. The success of these theories has only been partial because individuals do not always behave rationally. Although we are more often rational than not, our biases and limitations prevent us being fully rational.

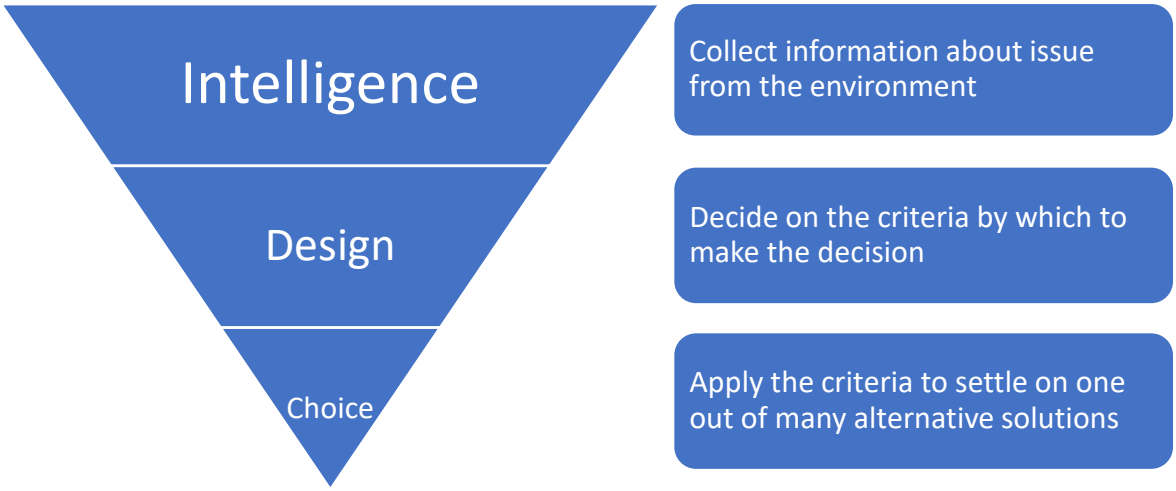
The history of humankind can be seen as a history of exploitation of one person by another. Not entirely, of course, but certainly an important element of this history is that we as humans are aware of each other's weaknesses and have an impulse to exploit them.

The exploitation of irrationality is evident in the digital world, which uses different factors to target and influence us in a variety of ways. There are many styles of influencing, none of which lead to success with any certainty. An agent [2] himself prefers to influence, and prefers to be influenced. It is one step in the journey of becoming more confident and successful in influencing scenarios and outsmarting peers. People create their own "subjective social reality" from their perception of the input. Root cause [3] analyses suggest it is unconscious biases.

What is unconscious bias, then? At any given moment, we receive over 11 million [4] bits of information from the world around us. Our brains can consciously process only 40 bits of information. To avoid overload, our brains use shortcuts, shaped by past experiences, cultural norms, and personal beliefs, to determine which information is important. These shortcuts are called unconscious biases, and there are hundreds of them operating outside of our awareness. If we do not implement intentional tactics and mechanisms to challenge our biases individually and as a group, they can negatively affect the quality of our individual and group decisions and relationships. If we cannot eliminate unconscious biases, they lead us to make irrational decisions. Herbert Simon's model [5] of decision-making suggests that decision-makers do not have complete information or unlimited cognitive resources at their disposal, which leads to a process of bounded rationality. This means that decision-makers must use heuristics, or mental shortcuts, to make decisions that are satisfactory rather than optimal.

According to Simon's model, the process of decision-making can be broken down into three stages: intelligence, design, and choice. In the intelligence stage, decision-makers gather and process information about the problem at hand. In the design stage, they generate and evaluate alternative

courses of action. In the choice stage, they select the most promising course of action and implement it.



2. Herbert Simon Model on Decision Making (1947)

Herbert clearly stated that decision-makers gather and process information about the problem to resolve it and after 2001 people mainly use internet to gather the information in U.S.A. The interaction with internet can be recorded and used to depict the collective human behavior. Specially periods of anxiety may precede trends to sell on the financial market at lower prices. People may tend to learn more about the state of the market during such concerning times and such interaction may produce warning signals about the economy and financial market.

Personal factors, as well as technical factors, are thought to play a major role in decision-making, as we may see in the context of this study, in which we focus on investor behavior and stock sentiment in the US market. The purpose of every investment should be to provide some kind of income or capital gain. People like to invest when things are certain for them and usually individual investment decisions are influenced by personal factors such as age, education, income, and investment portfolio. Their investment decisions are at the same time influenced by the complex financial models used to predict market movements.

The study of economics has greatly facilitated our understanding of finance and investing. When it comes to economic and financial decisions, traditional economics assumes that people are rational. People respond to incentives rationally because their goal is always to maximize benefits while minimizing costs. Although not everyone agrees on what constitutes a benefit and what constitutes a cost, there tends to be some consensus in a market with millions of participants. Information plays a significant role in investment decisions to minimize the risk associated with these decisions. More informed investors can make better decisions, so they try to look for more market information to lower the uncertainty in investment choices (Lin, 2002).

Behavioral finance is a new paradigm of finance that introduces behavioral components to decision-making. Behavioral finance contradicts the efficient market hypothesis [6] by attributing market inefficiencies to non-rational investors. A rational investor is one who (I) always updates

his beliefs when he gets new information in a timely and appropriate way and (II) always makes choices that are socially acceptable (Thaler, 2005). According to Nofsinger (2001) the field of finance has changed over the past few decades based on the idea that investor make rational decisions and may predict the price movements more accurately. There are several theories and models' investors use to evaluate stock market and its components when they make investment. We will cover them below.

Efficient Market Hypothesis

Economic and financial theories [1,2,6,7,8] assume that individuals make rational decisions by considering all available information. The emergence and evolution of efficient market theory [6], whose belief in rationality contributes to the idea of an efficient market. The efficiency of the market has been the focus of finance theory for the past five decades to enable market participants to develop valuation models, and finance theory [8] covers all market processes.

EMH's theoretical underpinning is comprised of three major ideas. (i) It is impossible to earn credible risk-adjusted profits based on historical data because investors are rational and value securities rationally. This is what is referred to as the "weak" form of efficiency, and it occurs when all previous market prices, returns, and other information are fully incorporated in prices. This renders technical analysis useless (ii) irrational investors' random transactions cancel out without impacting prices. The "semi-strong" variant says investors can't gain higher returns utilizing public knowledge since it's already in pricing. This renders fundamental analysis useless (iii) Arbitrageurs remove unreasonable investors' market impact. "Strong" EMH indicates that all public and private information is represented in security pricing.

Table 1.01 Efficient Market Hypothesis Summary.

Form of Efficiency	Definition	Implications
Weak	All previous market data is fully incorporated in prices, making technical analysis useless.	It is impossible to earn credible risk-adjusted profits based on historical data.
Semi-Strong	Public knowledge is already reflected in pricing, rendering fundamental analysis useless.	Investors can't gain higher returns utilizing public knowledge.
Strong	All public and private information is reflected in security pricing, making insider knowledge irrelevant.	Insider knowledge won't help investors earn more.

Even insider knowledge will not help investors earn more. Since it was difficult to accept the strong form, many evaluations were based on the weak and semi-strong forms. There was also evidence that insiders earned abnormal gains while trading lawfully [9,10]. Fama (1965) discovered that stock prices followed a random walk pattern [11,12]. Fama et al. (1969) conducted an event study to test the semi-strong form of the efficient market hypothesis. They analyzed the impact of unexpected corporate earnings announcements on stock prices. The study found that the stock prices reacted quickly and accurately to the news, suggesting that the market was efficient

and that all publicly available information was already reflected in stock prices. This supported the semi-strong form of the efficient market hypothesis. Over the last 40 years, the Efficient Market Hypothesis (EMH) has been a major financial paradigm, and it has also been one of the most heavily questioned one.

The efficient market hypothesis, as explained by Fama in 1970, informs that financial markets are efficient. Fama in his study argues that it is not possible to consistently make profits by trading based on current data. In the 1970s, the EMH gained significant popularity, and many studies were committed to prove its validity. These studies were highly successful both theoretically and empirically.

The fluctuations in the stock market may be influenced by human biases and cognitive errors, which is an area of study in behavioral finance [13]. Statma [14] explained that unlike the perfectly rational individuals assumed by traditional finance, "Real" people are used in behavioral finance. Specifically, Thaler [15] thought it was important to distinguish between the nonexistent completely rational investor and the more realistic quasi-rational investor. The quasi-rational investor makes reasonable attempts to invest wisely yet is vulnerable to common pitfalls. The semi-rational investor is an important consideration in the field of behavioral finance [16].

Modern Portfolio Theory (MPT)

Modern Portfolio Theory (MPT) [17], developed by Harry Markowitz in 1952, is a cornerstone of modern finance and has had a significant impact on investment management practices. The theory suggests that investors can optimize their investment portfolios by diversifying across multiple asset classes with varying levels of risk, rather than relying on a single asset or a few concentrated investments. Key concepts for MTP are diversification, expected returns, risk, efficient frontier, if we can summarize the above can reach out below list:

MPT is based on the following assumptions:

1. Investors are risk-averse, meaning they prefer to avoid risk if possible.
2. Investors can measure the risk of an asset or portfolio by its expected return and its variance.
3. The expected return of a portfolio is the weighted average of the expected returns of its constituent assets.
4. The variance of a portfolio is the weighted average of the variances of its constituent assets, plus the covariances between the assets.

As we can point out that in the MTP diversification is the process of spreading risk across different assets. By diversifying, investors can reduce the overall risk of their portfolios without sacrificing too much expected return. Shortly we can summarize the MTP with below pros cons table which it is useful for us at the end to compare with other traditional style models.

Table 1.02 Modern Portfolio Theory Summary.

Pros	Cons
Recommends diversification to reduce risk associated with individual investments.	Based on a number of assumptions, some of which may not be realistic. Assumes investors are risk-averse and rational, which may not be the case.
Is a relatively simple and easy-to-understand theory.	It is a static theory, and it does not account for the dynamic nature of the markets.
Provides a framework for optimizing the trade-off between risk and return.	Assumes normal distribution of returns and static correlations, which may not hold in reality.
Offers the concept of efficient frontier to show the optimal portfolio allocation based on risk tolerance.	May not account for black swan events or systemic risks.
Empirical evidence supports the effectiveness of MPT in improving portfolio performance and reducing risk.	Historical data used to estimate expected returns and risk may not be a reliable indicator of future performance.

Always historical performance cannot be good predictors in long horizon. Therefore, the investors always looking for the advanced financial models to progress their position in the financial market. Capital Asset Pricing model [18] can be perfect example to consider the risk and its return.

Capital Asset Pricing Model

CAPM [18] is the most well-known method for modeling stock returns (CAPM). The Capital Asset Pricing Model (CAPM) was developed independently by William Sharpe, Jack Treynor, John Lintner, and Jan Mossin in the early 1960s.The CAPM classifies drivers of stock returns as either systematic or idiosyncratic. Idiosyncratic risk is distinct to each investment and may be mitigated by portfolio diversification. Systematic risk, however, is connected to a security's return sensitivity to the market and cannot be eliminated by diversification, suggesting that investors are paid with better returns for assuming the higher risk. This is likely the simplest example of what would later be termed a fundamental factor, explaining the projected stock returns of an asset as a function of its market beta.

As above mentioned different financial models, such as the CAPM, can be used to obtain information that can be used to make investment choices.

CAPM Equation:

$$E(R_i) = R_f + \beta_i\left[E(R_m) - R_f\right]$$

Where:

$E(R_i)$represents the expected return on asset i

R_f represents the risk-free rate of return

$\beta_i$represents the beta of asset i

$E(R_m)$ represents the expected return on the market

The CAPM is widely used and accepted easy to apply financial model by the investors. However, there are other financial models which considered more advanced than MTP, EMH and CAPM it is Arbitrage Pricing Theory [19]. CAPM uses a single factor, which is the asset's correlation with the overall market, or its beta however ATP is multifactor complex model.

Arbitrage Pricing Theory

Stephen Ross [20] developed the model as an alternative for the Capital Asset Pricing Model (CAPM) in 1976. The name "factor" wasn't popularized until Ross introduced APT years later. The APT asserts that asset return variables may fluctuate over time and across markets. According to the APT, certain factors can contribute to excess returns over long periods, but they can also suffer significant volatility over shorter time frames, leading to underperformance.

The APT argues that the return of an asset is influenced by a number of variables unlike CAPM, including macroeconomic factors, industry factors, and firm-specific factors.

$$r_i = \beta_{i1} f_1 + \beta_{i2} f_2 + \beta_{i3} f_3 + \cdots + \beta_{ik} f_k + \epsilon_i$$

In above equation, r_i represents the return of asset i, β_{ik} represents the sensitivity of asset i to factor k, f_k represents the value of factor k, and ϵ_i represents the idiosyncratic error term for asset i. The model is a multifactor model, which seeks to explain asset returns based on multiple risk factors [21].

While certain factor exposures display excess risk-adjusted returns over extended periods of time, they may also endure severe cyclicality over shorter timeframes, including underperformance [20,21]. Along with market cycles, investment needs fluctuate over time. This can be due to past financial crises and market sell-offs, resulting in a focus on liquidity management, mixed hedge fund performance over the last decade, asset managers experiencing fee pressures and calls for increased transparency, and a preference for risk-based portfolio management over traditional allocations to generate long-term premiums [22]. In most cases, investors give more weight to a select few criteria that are founded on sound core notions that have been studied for decades.

Other non-traditional, alternative factors have lately piqued the interest of investors, who are attempting to integrate their classic factor-investing abilities with the prospects provided by information technology and big data. It is commonly understood that news information flows are driven by rapid dynamics [23] and are typically absorbed pretty quickly by markets. This is not to say that slower moving signals do not exist. However, they are certainly more concealed inside

the massive volumes of data. Interestingly, in more mature and liquid markets, news information gets absorbed and priced quickly.

There are other traditional financial models however we only mentioned building block of traditional financial models, nowadays investors' demand is more peculiar risk assessments for their investment preferences therefore behavioral finance got much attentions since it can explain biases and pitfalls which lead to huge fluctuations in the stock market. Traditional financial models assume that investors are rational and make decisions based on sound economic principles as we drew above models.

This work is assumed to provide an approach for combining news sentiment signals [24] with fundamental components in order to create multi-factor portfolio with minimal turnover and no bias which will be proxy for sentiment factor. We demonstrate this by building a basic news-based Sentiment Factor [26], but the same method may be used to other factors that depend on alternative data sources.

Incorporating a sentiment factor in the APT multifactor model would involve adding a new variable to the equation that captures the impact of news sentiment on asset returns. This would require estimating the beta coefficient of the sentiment factor (referred to as the $\beta_s entiment$)

Adding sentiment to the ATP equation:

$$r_i = \beta_{i1} f_1 + \beta_{i2} f_2 + \beta_{i3} f_3 + \cdots + \beta_{ik} f_k + \beta_s Sentiment * SentimentFactor + \varepsilon_i$$ [12]

Here, the sentiment factor represents the sentiment factor for each asset and $\beta_s Setiment$ represents the sensitivity of each asset's returns to changes in the sentiment factor. There are several studies that have incorporated sentiment factors int the APT model. For example, a study by Huang and Zhou (2019) used a combination of news sentiment and other factors to build a multifactor model for asset returns. In another study by Wang et al. (2020), sentiment factors derived from social media data were incorporated into the model APT. Let's summarize the traditional financial models that we can use in our estimation of the thesis hypothesis with their pros and cons to have clear picture what and why the model is suitable for us for sentiment factor estimation. The **Table 1.03** summarizes different models used in finance to estimate the expected return on equity. The Capital Asset Pricing Model (CAPM) is a widely used liner model that calculates expected return based on systematic risk and beta estimates, but it has limitations in its applicability and ignores unsystematic risk. The Arbitrage Pricing Model (APM) is more flexible and captures idiosyncratic risk, but is limited by uncertainty in estimates and unnamed factors. The multifactor model is more comprehensive and intuitive than the APM, but it can be more complex.

[12] The equation will be explained in more details in empirical part.

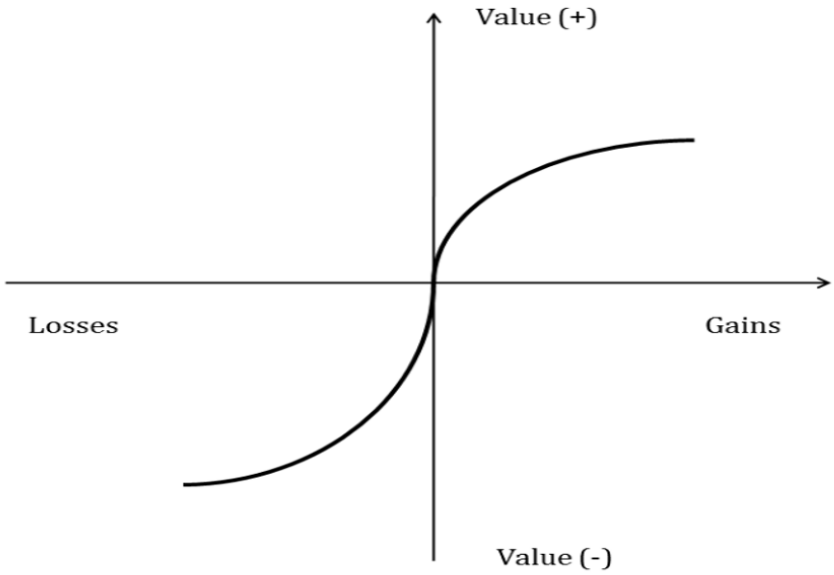
Table 1.03 Traditional Financial Models Summary.

Model	Expected Return on Equity	Pros	Cones
The CAPM	$E(R) = R_f + \beta \times ERP$	1. Simplicity 2. Systematic risk 3. Widely used	1. Assumptions based 2. Limited applicability 3. Beta estimates 4. Ignores unsystematic risk:
APM	$E(R) = R_f + \sum_{j=1}^{n} \beta_j (E(R_j) - R_f)$	1. Additional factors 2. Flexibility 3. Captures idiosyncratic risk	1. Uncertainty of estimates 2. Factors are statistical 3. Factors are unnamed 4. Data availability
Multifactor model	$E(R) = R_f + \sum_{k=1}^{n} \beta_k (E(R_k) - R_f)$	1. More comprehensive 2. Greater flexibility 3. Captures idiosyncratic risk 4. More intuitive than APM 5. Better prediction	1. Factors are unstable 2. Complexity
Alternative models	AI and ML powered	Can eliminate the above cones using factor-based mode.	Parameters for the models might be difficult to estimate and fit
Proxy models	$E(R) = a + bX_1 + cX_2$	1. Easy to interpret 2. Easy to update	1. Variables in model may not be proxies for risk 2. Susceptible to error

The traditional models are the widely used fundamental stock valuation techniques. However, decision making process is a complex process and to understand human biases and avoid them or detect in the market can be pivotal change for the investor therefore it worth to take a look behavioral finance concept.

Behavioral finance

Behavioral finance [1,2] explores how emotional and cognitive biases affect investment decisions. Understanding cognitive biases can help achieving client goals, portfolio construction, and discovering investment choice discrepancies. Behavioral finance explains market abnormalities. The interpretations suggest that behavioral and traditional finance may work better together than separately. Behavioral finance is a developing field related with the social influence on decision-making when people are faced with uncertainty. Leveraging methodologies derived from artificial intelligence (AI), I evaluated the degree to which specific behavioral finance concepts like overconfidence, herding, the gambler's fallacy, and mental accounting elucidate individual investor decision-making processes within the United States market.



2. Prospect Theory Value Function, Source: Kahneman and Tversky (1979)

Prospect Theory delineates framing (editing) and evaluating risk. Tversky and Kahneman's Prospect Theory explains how people handle risk and uncertainty. This theoretical framework explains the apparent inconsistency of individuals behavior when assessing risk within ambiguous circumstances. It contends that individuals are not invariably risk-averse; they exhibit risk aversion for gains but risk seeking for losses. Due to the certainty effect, certain outcomes loom larger than probable ones. On the other hand, framing influences choices by altering the reference point for evaluating outcomes.

Framing refers to the multiple ways in which the same problem is phrased and presented to decision makers, and the effect examines how framing might influence decisions in a way that is inconsistent with the traditional axioms of rational choice. Systematic preference reversals were also observed when the same problem was presented in several formats (Tversky and Kahneman, 1981). In Prospect Theory and Modern Portfolio Theory, the function for maximizing value is different. In the modern portfolio theory, the goal is to have the more gains at the end, while the prospect theory looks at both gains and losses. This is because different people may make different decisions in situations where they end up with the same amount of money, such as the endowment effect and the sunk cost fallacy.

Behavioral biases [27] may cause cognitive mistakes in investors, below listed the well-educated and researched biases.

Behavioral Biases

Overconfidence Bias:

Overconfidence bias [28] is the propensity for people to place more faith in their own evaluations or conclusions than is supported by the facts. Because of this tendency, people may exaggerate their knowledge, understate the dangers of a circumstance, or make too optimistic predictions. A sentiment index, which gauges the general positive or negative sentiment of a financial market. A Sentiment index could be affected by overconfidence bias if people tend to express more optimistic or positive opinions than are supported by the data because they are overconfident in their own judgments or talents. However, it is a challenging task for us to make quantitate analysis to measure the news sentiment considering overconfidence bias.

To take the closer look we need to deep dive the idea of overconfidence, which originates from a huge corpus of cognitive psychology studies and surveys in which participants overestimate both their own prediction ability and the accuracy of the information they have been given. It can be carried out through surveys to capture such effects. On the other hand, people are terrible at assessing probabilities; occurrences that people believe are certain to occur are often significantly less likely to occur than they believe. In summary, individuals mistakenly believe they are more intelligent and have more knowledge than they really do (Pompian, 2006) [29].

Some studies have been well documented by Fagerström (2008) [30], he studied overconfidence and over positivity in the market and how they affect investment decisions. The study uses a quantitative back-testing approach based on IBES (Institutional Brokers Estimate System) data. The data provides a summary of predicted profit growth for S&P500 firms for the next 12 months vs. actual growth from February 1986 to April 1987. The results of the study indicated that analysts covering the S&P 500 companies were both overconfident and overoptimistic in their earnings forecasts. Specifically, the forecasts were found to be too high compared to the actual earnings growth, suggesting that analysts were not able to fully account for the risk and uncertainty inherent in their predictions.

Representativeness Bias:

Representativeness bias arises when people assess the likelihood of an event depending on how closely it matches a typical or representative example of a certain category or group. In the financial markets, representativeness bias can cause investors to make decisions based on their opinion of how typical or representative a specific investment is, rather than its real risk and return characteristics.

In the field of behavioral finance, representativeness is referred to the concept of conditional probabilities. It implies how people frequently mistakenly evaluate the likelihood of an event-based likelihood. Using the heuristic, you can figure out how likely it is that an object or event A is part of a class or process B. People use representativeness most of the time, according to Tversky and Kahneman (1983) [1], when they are asked to figure out how likely it is that A belongs to B. Representativeness can be shortened to "similarity" if A and B are described in the same way.

For example, making decisions based on stereotypes is an example of representativeness. Investors tend to base their selections on what they've seen or heard in the past. (Shefrin, 2000). Ritter (1991) pointed out another interesting effect of the representativeness bias, which is that IPOs don't do well in the long run because investors are focused on the short term. This has many effects on how investments are chosen. When it comes to investing, most people believe that positive factors about a company are instantly bound to positive factors about its stock. Most of the time, these companies are not viable investments (Lakonishok et al, 1994). Stock market sentiment and representativeness bias, as both can be influenced by people's beliefs and prejudices about specific investments or industries. Overall, representativeness bias can lead investors to make decisions that are not based on a complete study of an investment's risks and prospective returns, which can have a detrimental impact on their portfolio. It is critical for investors to be aware of this bias and avoid allowing it to impact their investment decisions.

Herding Bias:

Herding bias [53] refers to people's bias to mimic the behaviors or decisions of others rather than making their own independent decisions. This can happen in the stock market when people follow the buying or selling decisions of other investors rather than making their own judgments based on a thorough study of an investment's risks and possible returns. The core premise is that under certain conditions, people might ignore their private information and follow the crowd, resulting in herding behavior.

In the context of the financial markets [54], the term "herding" implies the technique of mutual imitation, which eventually results in activity of convergence. This is the most frequent error committed by investors when they align their investment choices with the majority's consensus. When the best time to buy or sell is near in the financial markets, even the investor who thinks he should act, feels a strong psychological pressure to not do so. This is due to the weight of public information becomes so strong and which can create clear trend or fluctuation which in turn overwhelms any individual's private information. For example, in 2008, when Reliance Power had its initial public offering (IPO), many investors bought shares even though they didn't know everything about the company. Investors use "herd behavior" because they want to know what other people think of their choices (Scharfstein and Stein, 1990).

Economou, Kostakis, and Philippas (2010) used daily data from the Greek, Italian, Portuguese, and Spanish stock markets from 1998 to 2008 to examine herd behavior in extreme market conditions. They also analyzed herd behavior during the 2008 financial crisis. According to the findings, herding is more prevalent in rising stock markets, leading to bubbles.

Let’s suppose there's a significant increase in positive posts about a specific stock on social media. This surge in attention, combined with an uptick in buying activity for that stock, could indicate herding behavior driven by positive sentiment. By tracking these patterns over time and across various stocks or markets can uncover the herding behaviors in the US stock.

To factor in herding bias in a sentiment index, one approach could be to analyze the volume or frequency of buying or selling activity alongside sentiment variables for a particular stock or portfolios on the newswire. Other relevant data that could be considered could include the volume or frequency of media coverage or social media posts about a particular stock or portfolios, as this could provide insight into the level of attention or interest it is receiving from investors.[55]

Anchoring Bias:

Anchoring bias in finance refers to the tendency of people to rely too heavily on the first piece of information they receive when making decisions, even if that information is not relevant or accurate. This bias can lead people to make irrational financial decisions, such as overpaying for a stock or undervaluing an investment opportunity.

Investors make irrational investment decisions when they place an excessive value on psychologically motivated and statistically random "anchors." Should I purchase the stock now or wait for a better opportunity?' or 'is the stock fairly priced? If investor looking for the answer of those questions by random and psychologically driven 'anchors,'- initial value which producing results that are biased.

Kristensen and Gaerling (1997) examined the hypothesis that "in negotiations counteroffers are created via an Anchoring-and-adjustment process leading to an effect of the anchor point, and those counteroffers are impacted by changes in reference point" In a simulation of the negotiation process, undergraduate students studying business discovered that participants utilized the proposed selling price as an anchor in their negotiations.

To mitigate the effects of anchoring bias, it is essential for investors to carefully consider all available information and consider multiple viewpoints before making a financial decision. It is also helpful to seek the advice of a financial professional or to use tools such as financial modeling to help make more informed decisions.

Gamblers' Fallacy Bias:

Gamblers' fallacy bias, also known as the "Monte Carlo fallacy" or the "fallacy of the maturity of chances," is a cognitive bias that occurs when people believe that a certain outcome is more or less likely to occur based on past events. This bias is often seen in gambling, where people may believe that a certain outcome is due to happen based on the outcomes of previous plays or spins.

According to Kahneman and Tversky (1971), gambler's fallacy stems from an erroneous belief in the impartiality of chance. This predilection, referred to as the Gamblers' Fallacy, manifests when investors inaccurately anticipate a reversal in a given trend, leading to an inclination towards contrarian reasoning. The Gamblers' Fallacy is characterized by the investor's tendency to act under the assumption that anomalies in random sequences are intrinsically self-correcting. To illustrate, if a fair coin is consecutively flipped 10 times, landing on heads each instance, an investor who anticipates the subsequent flip to result in tails is understood to be under the influence of this bias. This cognitive error is demonstrably prevalent when an investor operates on the premise that deviations in stochastic events inherently self-correct.

Availability bias:

In behavioral finance, availability bias refers to people's inclination to base their decisions on information that is readily available to them rather than evaluating all relevant information. This bias can lead to irrational financial decisions since people may not have access to all of the information they need to make an informed decision. Availability bias was first identified and

studied by psychologists Amos Tversky and Daniel Kahneman in the 1970s. They observed that people tend to rely more heavily on information that is easily accessible in their memory, rather than considering all of the relevant information. This bias can lead people to make irrational decisions, as they may not have access to all of the information that they need to make an informed choice. Before making a financial decision, investors should gather as much information as they can in order to lessen the effects of availability bias.

Loss Aversion Bias:

Framing bias is a behavioral finance bias that refers to the tendency to be influenced by the way in which information is presented or framed. This means that people can be swayed by the way in which a problem or decision is presented, even if the content of the problem or decision is the same.

As described above, behavioral finance biases examine stock market investor behavior in the context of uncertainty. The study of anomalies by economists and financial experts who aimed to comprehend the causes of irrational occurrences brought progress to this area.

Recent democratization (ease access of retail investors to the financial market) and COVID-19 [31] lockdowns increased financial market volatility. Market sentiment has become increasingly negative. (The term market sentiment refers to investor attitudes or general outlooks, or more specifically, how they feel about a particular investment or the entire financial market.)

During the COVID-19 lockdowns companies that constitute the S&P500 index showed signs of a sharp slowdown and a decline in earnings [32]. The scale of this event also brought about the halt of dividend payments and increased bankruptcy filings in the US. In such environment Investors' behavior and emotional mood swings cannot be explained in the framework of modern financial theory: the concept of behavioral finance is needed to understand the link between human psychology and market reactions. This study's analysis is centered on sentiment analysis, which is both textual and behavioral in nature. To understand the textual analysis, we need to dive Artificial Intelligence techniques.

Artificial Intelligence

Artificial Intelligence (AI) techniques are increasingly being utilized in the financial sector, particularly in areas such as asset management, algorithmic trading, credit underwriting, and blockchain-based finance [33,34,35,36]. This is made possible by the abundance of data that is currently available as well as the affordable computing capacity of computers.

AI techniques is enhancing risk-adjusted returns from trading or investing, managing event risk when investing or market-making, assisting compliance and surveillance analysts, and promoting trading activity. The technology can also help to improve risk management and promote better trading execution. Artificial intelligence is a proxy for our sentiment analysis. To carry out the sentiment analysis I have utilized AI's machine learning methods.

Machine Learning and Sentiment Analysis

Machine learning (ML) [37] is a type of artificial intelligence that allows models to automatically improve its performance and predictability through experience and training data, without need for much programming. ML models make use of large amounts of data in order to learn and improve their predictability and performance. In addition, newly developed AI techniques can record and collect data from online sources and analyze collective irrational and rational human behaviors [36]. Generally, to optimize operational workflows and risk management, as well as to detect signals and identify hidden patterns in big data, ML models are essential [38]. Sentiment analysis is measure of extracting and quantifying subjective information from text data using techniques from Natural Language Processing (NLP) and ML.

The financial market has become more complex and technology-oriented, stock market sentiment has become a key determinant of large market movements. As Tetlock's [40] research showed that news stories contain information that can be used to predict both earnings and stock returns. This has significant implications for the financial industry, as it highlights the importance of considering news sentiment in investment decisions. Bollen et al. (2011), Yu et al. (2013), Smailović et al. (2014), and Walker (2016) [44], have also examined the impact of social media on stock market volatility. They found that social media platforms such as Twitter can be used to predict changes in stock prices and volatility. Missing important information that has an impact on investor position or portfolio. In this context, it is worth looking into the role of emotions such as fear, greed, and excitement in influencing investment decisions.

Sentiment analysis can be done by scraping or crawling newspapers and social networks. We will avoid traditional methods (Baker, M., & Wurgler, J. (2013)) [40,41,42] such as surveys dividing investors into clusters. Because NLP enables the transformation of massive volumes of unstructured textual information. Such as news, into simple machine-readable insights that may be utilized as signals to make rational decision. NLP and ML algorithms are able to analyze millions of news articles and social media posts. The ML model converts this information into actionable insights that can be used to identify trends and opportunities in the financial markets in the concept of behavioral finance [39].

The use of sentiment analysis can range from simple content aggregation methods to more advanced machine learning techniques that attempt to uncover the complex relationship between different events in the media. By analyzing the tone and content of news articles and social media posts, sentiment analysis can provide information on market sentiment, which can impact stock prices and returns [42].

AI and machine learning can digest large amounts of data and draw precise conclusions and make accurate predictions, the question arises whether Artificial Intelligence based sentiment index is a rational tool for investors. To measure it we need to create the sentiment-based factor that can be informative tool for the stock market.

Data Description and Methodology

Sentiment analysis of news and social media can help investors gain insight into the market and make more informed decisions [41]. By analyzing a tone and content of news articles and social media posts, sentiment analysis can provide information on market sentiment, which can impact stock prices and returns. To conceptualize it for financial market Dow Jones, Reuters, MT Newswires, Alliance News, FX Street, The Fly newswires[13] are used sources of text based financial and economic news and information. They provide real-time updates on various financial markets, as well as regulatory news and press releases from a variety of sources, including 21,000 web publications. These sources are used by financial professionals and individual investors alike to stay up-to-date on the latest market developments and to make informed investment decisions. By aggregating and analyzing the vast amount of information produced by these newswires, investors can gain valuable insights into market trends and conditions, and use this information to adjust their investment strategies**.**

To aggregate the newswire data, I utilized Python libraries and its state of art environment. Using Python programing language, I can have access to real-time information in the global financial markets. To construct a sentiment factor as well as to compare output with fundamental data driven MSCI Barra GEM3 model [45], the above-mentioned newswires and websites [14]were used respectively as the primary data source. The emotional features of the text being extracted as the initial step in the sentiment analysis process. The acquired data analysed and weighted according to their importance and relevance in the concept of BF. The sentiment data then compiled and used to calculate the overall sentiment factor and sentiment trends in the S&P 500 stock market[15].

The NYSE's SPX500 is comprised of 500 constituents [43], each represented by a unique ticker symbol. These ticker symbols are identifiers for the listed companies. In this study, the data used for analysis was randomly selected based on ticker from a sample of 500 companies. Randomly selected companies' ticker is 250 out of 500. Additionally, respective external data was downloaded, mainly including the Fama-French 5 factors, momentum, and traded liquidity. The period is Jan 1st, 2012 to July 1st, 2022 both for traditional factors and sentiment-based factor. Our aim is to understand the impact of these factors on the performance of the selected stocks and then open the new dimension to compare with synthetic sentiment factor which empowered by behavioural finance elements.

To construct an alternative sentiment-driven factor based on news sentiment scores and fundamental data, I need to scrap the newspapers and websites. Although there are several ways

[13]See Dow Jones: https://www.dowjones.com/
Reuters: https://www.reuters.com/
MT Newswires: https://www.mtnewswires.com/
Alliance News: https://www.alliancenews.com/
FX Street: https://www.fxstreet.com/
The Fly: https://thefly.com/

[14] See h*ttps://finance.yahoo.com and https://www.macrotrends.net/*

[15] See S&P 500 index is composed of 500 large-cap stocks from leading companies in various industries in the United States. The list of selected stocks is included in Appendix I, along with information on the sources used to obtain the desired data.

to collect this data, such as manually searching for articles on news websites, but these methods can be time-consuming and may not capture a comprehensive set of articles or posts. On the other hand, I can use public APIs (Application Programming Interfaces) which is costly. Therefore, web scraping[16] tools was opted by me to automate the data gathering process and collect a larger and more diverse set of articles.

To the tickers incorporated respective company names [43] and out of these combinations were generated URLs for each ticker to iterate through every available text in the newswires for the period of daily, weekly and monthly bases. The data obtained from raw text extraction was used to arrange and categorize the text into a set of predefined titles based on its sources. As demonstrated the below table.

Table 1.04 Predefined Titles for Extracted Text.

Predefined categories	Description
Fact	This value indicates that the detected event is based on a concrete statement of information over ***company sources***.
Forecast	This value indicates that the detected event provides ***guidance or predictions*** about the future.
Opinion	This value indicates that the detected event expresses a ***view or a hypothesis*** about a statement of information.
Mention	This value indicates that the detected event is likely more about or related to the general topic, ***without providing a specific factual statement.***

Above **Table 1.04** summarizes information on the degree of relationship between the detected event and a factual statement. The text has four possible values: "Fact", "Forecast","Opinion", and "Mention". Each value describes the type of event detected, with "Fact" indicating a concrete statement of information, "Forecast" indicating guidance or predictions about the future, "Opinion" indicating a view or hypothesis about a statement of information, and "Mention" indicating an event that is likely more about or related to the general topic. I have developed a proprietary technological solution that categorizes documents into a taxonomy of topics which are fundamental to today's business and thesis requirement. Methods of categorization that are derived from linguistic research WordEmbeddings[17][46]. Deep learning architectures have attracted

[16] Web scraping tools include BeautifulSoup, Scrapy, and Selenium, which can be utilized to extract data from websites in various formats such as HTML, XML, or JSON.

[17] Word Embeddings are the texts converted into numbers and there may be different numerical representations of the same text. (https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/)

considerable attention in NLP due to their state-of-the-art performance on various NLP tasks. However, these architectures are inherently incapable of processing textual data in its raw form. Instead, numerical data is required as input to these architectures to perform tasks. As mentioned cleaning up the data is essential for our sentiment analysis. We have successfully removed stop words and then we can tokenize the text. Tokens or values that have been compiled into a language are converted to "csv" file format and excel templates. Tokens could be numbers or dates. Token may be reduced to text's most fundamental form, or they can be used exclusively for a certain tense; either way, outcome will have a Parts of Speech ([18]PoS) signature [48]. In NLP, PoS tagging assigns each word in a sentence a part-of-speech. This is used in information retrieval, text-to-speech conversion, and word sense disambiguation. There are standardized sets of tags, like UPenn TreeBank II. In addition, the database encompasses a plethora of supplementary data including temporal information, such as the time and date of the story, as well as its headline, source, genre, and references to other related news items. Splited(Usual distribution: 60% training, 20% validation and 20% test) data into three sets: training set, validation set, and testing set. The training set used to train the model, the validation set used to tune the model hyperparameters, and the testing set used to evaluate the model's performance. To deploy a machine learning model used a cloud-based service, Amazon Web Services. The RNN [19]was opted for the model to tune the data with the following process in the Amazon Web Service cloud infrastructure: Initialize the model parameters, Define the loss function, Choose an optimizer.

The model described above is able to identify the entities mentioned in a news article, extract the subject or type of event, determine the role of each entity within each event, and consolidate the information. The model then uses the context in which the entities are mentioned to determine the role of each entity within each event. Finally, the system consolidates the information by combining multiple recognitions or inferences of entities and events within the same document to determine the final event category for the entity(ies).

After having filtered data according to title then I can delegate further clustering of the text according to Table **1.05** classifiers. The occurred events can be narrowed then to an entity level. It

[18]Pyhton Packages:
import nltk
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import spacy
Example for tokenize the sentence:
sentence = " Charles University is the best."
tokens = nltk.word_tokenize(sentence) ,This can be an entire word, a part of a word (called a subword or n-gram), or even punctuation marks.
[19] RNNs are well-suited for sequential data, such as text or speech, however CNNs are well-suited for spatial data, such as images.

would allow us to break down the analysis into sentiment bins and then elaborate it with behavioral finance concept.

Table 1.05 Event Classification Structure (*Sample Data*)**.**

Topic	Group	Type	Sub-type	Role	Category
Stock Market	Company Earnings	Financial Events	Quarterly Earnings Report	Shareholder, Analyst	Company Name, Industry
Stock Market	IPOs	Corporate Actions	Initial Public Offering	Investor, Underwriter	Company Name, Industry
Stock Market	Mergers & Acquisitions	Corporate Actions	Acquisition	Acquirer, Target, Adviser	Industry, Deal Size
Stock Market	Economic Data	Macroeconomic Events	Employment Report	Investor, Economist	Data Type, Time Period
Stock Market	Market Volatility	Market Events	Stock Market Crash	Trader, Investor	Time Period, Magnitude

In **Table 1.05** the header of table is classifiers however the provided respective data is a sample and the lexicon used here is ***confidential data***. Event extraction from text involves the segmentation of text into smaller units such as words, sentences, and parts-of-speech. This process is followed by the application of semantic attributes and predefined criteria to identify document events. The contextual information surrounding each entity is considered to determine its role in a particular occurrence. In some cases, multiple entities may be allocated roles in an event, such as a respondent in a legal event. Finally, the detected or inferred entities and occurrences within the same document are consolidated, and the final event category for each entity is assigned. Each criteria documented in Appendix I.

- Sentences with a similar amount of probabilities in the extreme positive and negatives will produce a neutral average score with a low confidence.
- In sentences with significantly more positive than negative probabilities, the average score will be positive and it will have a high confidence.
- For sentences that only have negative probabilities our model will give a negative average score with a very high confidence.

Text extraction, content analysis, and sentiment calculation are typically performed as a three-step process in our sentiment analysis study, once the process was finished then I created a custom dataset involves applying filters [C] or transformations to raw data to extract the desired information. The output dataset can be downloaded as a flat file in either "csv"or Excel format for further analysis or use. We compile the sentiment scores of news articles on a weekly basis for each financial institution under consideration. During this process, we consider the relevance score of each article and assign greater weights to those that are more innovative in nature.

The model generates scores and tags for each ticker to measure the sentiment, novelty, and prominence of the event related to the entity. The Event Sentiment Score [20]ranges from -1 to 1, with a score of -1 indicating negative sentiment and a score of 1 indicating positive sentiment. The Event Similarity Days is a score between 0 and 365, where a higher score suggests a more unique event. The Event Relevance Score ranges from 0 to 100, with higher scores indicating more importance of the event in the story.

The **Table 1.06** classifiers have been designed and trained to automatically categorize different types of financial content such as news articles and press releases into relevant themes or topics. The application of these classifiers enables the processing of vast amounts of unstructured data in real-time, resulting in high-quality structured data that can be utilized for a range of purposes.

Table 1.06 Sentiment Clusters.

Document_Sentiment	The overall sentiment of the document.
Entity_Sentiment	The sentiment score of the detected entity.
Entity_Text_Sentiment	The sentiment score of the text associated with the detected entity.
Event_Sentiment	The sentiment of the event being described in the document.
Event_Similarity_Days	The number of days used to calculate event similarity.

To exemplify this concept, **Table 1.07** portrays the combined weekly sentiment scores for 3M Co. from 2019/01/01 to 2022/01/01, and demonstrates that the sentiment scores are relatively erratic on daily/week/monthly bases. Consequently, we employ a filtering technique for numerically even we can apply the formulas which fields that extracts a significant signal from the turbulent data.

Table 1.07 Final Sentiment Scores According to Companies.

Entity_ID	Ticker_ID	Entity_Name	Entity_Sentiment_D_AVG	Entity_Sentiment_W_AVG_SENT	Entity_Sentiment_M_AVG_SENT	Entity_Sentiment_NEWS_COUNT	HEARDING_BEHAVIOUR
543900	CPB	Campbell Soup Co.	0.04832	-0.05570	0.15871	172	5.35811
619882	JPM	JPMorgan Chase & Co.	0.06644	0.06583	0.18601	171	1.55916

[20] Sentiment scores are a metric for measuring text sentiment. Scores can range from 0-100, where 100 is the most positive possible outcome and 0 is the least. ***For implementation please visit following website: https://livevox.com/glossary/sentiment-score/#gref***

Table 1.08 The Explanations of the Table 1.06 Variables.

Entity_ID	Randomly Generated unique number specifc for the company
Entity_Name	Legal registred Name of the entity
Entity_Sentiment_D_AVG_SENT	Average sentiment of the event occurred in **Daily** bases regarding specific company
Entity_Sentiment_W_AVG_SENT	Average sentiment of the event occurred in **Weekly** bases regarding specific company
Entity_Sentiment_M_AVG_SENT	Average sentiment of the event occurred in **Monthly** bases regarding specific company
Entity_Sentiment_NEWS_COUNT	Count of News (events happned for the filtered period) regarding specific company

Given that news items are disseminated on a constant basis, including weekends, holidays, and outside of traditional trading hours, it is crucial to compute sentiment-driven indicators shortly before the planned trading moment. This applies to all strategies that leverage sentiment, irrespective of the frequency of trading. In our study, we update our sentiment analysis on a weekly basis, and have selected the end of business on Mondays as the cutoff point for the news feed. This timing is also utilized for the calculation of sentiment and the execution of trades.

Table 1.09 Bias Categories of Behavioral Finance.

Bias Category of BF	Lexicons
Overconfidence	conservative, cautious, measured, realistic, prudent
Representativeness	evidence-based, rational, impartial, objective, unbiased
Herding	independent, individual, self-directed, uncorrelated, diversified
Anchoring	adaptive, flexible, open-minded, agile, responsive
Regret Aversion	objective, impartial, disciplined, systematic, unemotional
Gamblers' Fallacy	prudent, disciplined, methodical, systematic, data-driven
Mental Accounting	analytical, systematic, logical, objective, data-driven
Hindsight	future-oriented, forward-looking, predictive, anticipatory, proactive

An index that aims to measure overconfidence considers data sources that reflect investor confidence, such as the number of bullish statements made by analysts on the news articles, since we can measure the sentiment level and its tone, or the proportion of news articles that contain optimistic language. An index that aims to measure herding behavior considers data sources that reflect group behavior, such as the volume of news release and read rate related to a particular stock or the extent to which investors are following the actions of others.

I have factorized Behavioral Finance biases using the **Table 1.09** lexicons in regard to data set and to measure the level of risk associated with human factors in other words to identify patterns in investor behavior that may be influenced by behavioral biases.

The model is trained on a large number of news articles, and the sentiment of each article is

determined using event sentiment analytics. To avoid biases, the articles are selected from a range of years and sources. Each model is thoroughly evaluated to ensure it is accurate and has a similar error rate. The models yield 41 sentiment bins, with 21 being neutral, allowing for a more precise measurement of sentiment. These sentiment bins are used to create factors for behavioral finance, and a bias-free portfolio is constructed and evaluated against fundamentally factored portfolios.

Table 2.00 Sentiment Data Descriptions.

Topic	Information
Data sources	Content from over 30 web publications
Data format	Comma separated values (.csv) files
Data fields	84 fields
Entity identification fields	Entity_ID

The process of constructing the factor involves using sentiment signals based on the aggregation of sentiment scores from non-neutral, relevant, and novel news events daily. This sentiment signal is smoothed using moving averages over different time periods, ranging from 1 week to 90 days, to create independent sentiment factors. *Companies with higher sentiment scores are typically assigned larger positive allocations, on the other hand those with lower sentiment scores are given larger negative allocations.*

The aforementioned statement delineates the methodology used to construct sentiment-based factors for constituents of the S&P 500 index, utilizing a factor model based on the framework proposed by Fama and French. This approach comprises two key steps: the first step involves selecting pertinent factors for comparative analysis, while the second step involves generating the respective values for these factors. For the enhancement and to check the hypnotizes of this study Fama and French model extended to MSCI Barra GEM3 model [45].[21]

Traditional Factors[22]:

Value: This factor is based on the idea that stocks that are trading at a lower price-to-earnings ratio, price-to-book ratio, or other valuation metrics, have the potential to generate higher returns in the long-term.

Size: This factor is based on the idea that smaller companies have the potential to generate higher returns than larger companies.

[21] There are usually material differences between simulated or back tested performance outputs and actual results. The analysis and observations in this report are limited solely to the period of the simulated historical data.
[22] You can get more info from following link: https://www.msci.com/documents/10199/71ad36ee-cefe-4bca-92d8-b52202167031

Momentum: This factor is based on the idea that stocks that have performed well in the past are more likely to continue to perform well in the future.

Low Volatility: This factor is based on the idea that stocks that have lower volatility have the potential to generate higher returns with lower risk.

The factor model is a useful tool in finance and investment analysis that allows investors to understand the drivers of an investment's returns. Intermittent periods of underperformance are characteristic of traditional investment factors, and in some cases, these factors exhibit pronounced cyclical behavior. By breaking down the expected return of an investment into the expected returns of underlying factors, the factor model provides a more nuanced and complete picture of the investment's risk-return profile. This information can then be used to make more informed investment decisions, such as allocating assets more efficiently or reducing exposure to certain risk factors.

To construct multi-factor portfolios with a high degree of accuracy and reliability, it is essential to obtain data from sources that are recognized for their credibility and dependability. As per the guidance informed, the website "macrotrends.com" has been designated as the primary source for extracting the required traditional data that will be utilized in building the portfolios.

For example, MSCI Barra, a notable entity in the realm of financial analytics, undertakes the construction and regular modification of its factor models by utilizing a network of 56 data suppliers provides over 100 data feeds which is costly for me to accomplish with the same way.

To construct the value factor, sorted stocks based on their book-to-market ratio (B/M[23]) because stocks with high book-to-market ratios are typically considered to be value stocks, while stocks with low book-to-market ratios are typically considered to be growth stocks. Therefore, we need to have book value and market capitalization of the randomly selected companies.

The whole codes for data gatering developed inside of python Class method. The below function is part of the class method and named "scrape_data" which takes a single argument "ticker". The function uses the input "ticker" to construct a URL to scrape financial data from the website "macrotrends.net". The data is obtained from a table that displays a company's stock price, book value, and price-to-book ratio over time. The function uses the "requests" module to send an HTTP GET request to the website, and the "BeautifulSoup" module to parse the HTML response. The function then extracts the data from the table and stores it in a data frame called "df". This list is then used to create a pandas DataFrame with four columns: "date", "stock_price", "book_value", and "price_to_book_ratio". The function finally adds a column named "Company" to the DataFrame which contains the input "ticker" value. The function returns the DataFrame as output. #

[23] B/M ratio = Book Value per Share (B) / Market Value per Share (M)

```
def get_book_value(self):
    url = f"https://www.macrotrends.net/stocks/charts/{self.ticker}/{self.ticker}/price-book"
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')
    table = soup.find("table", {"class": "table"})
    data = []
    columns = ['date', 'stock_price', 'book_value', 'price_to_book_ratio']
    for row in table.find_all('tr'):
        cells = row.find_all('td')
        if len(cells) == 4:
            date = cells[0].text
            stock_price = cells[1].text
            book_value = cells[2].text
            price_to_book_ratio = cells[3].text
            data.append([date, stock_price, book_value, price_to_book_ratio])
    df = pd.DataFrame(data, columns=columns)
    df['Company'] = self.ticker
    return df
```

After having the "df" data frame which is time series data, can be used further analysis to calcualte the B/M ratio, in order to find B/M ratio for each ticker and store it csv file, for that I have developed below python function.

```
    @staticmethod
    def calculate_book_to_market_ratio(book_value_data):
        """ Calculates the book-to-market ratio for each row in the book_value_data DataFrame. Args:
            book_value_data (pd.DataFrame): A DataFrame containing 'book_value', 'stock_price', and 'Company' columns.
        Returns: pd.DataFrame: A DataFrame with an additional 'book_to_market_ratio' column. """
        # Create a deep copy of the original DataFrame
        book_value_data_BM = book_value_data.copy()
        # Remove dollar signs, commas, and convert the 'book_value' column to a numeric type
        book_value_data_BM['book_value'] = pd.to_numeric(book_value_data_BM['book_value'].str.replace('$', '').str.replace(',', ''))
        # Remove commas and convert the 'stock_price' column to a numeric type
        book_value_data_BM['stock_price'] = pd.to_numeric(book_value_data_BM['stock_price'].str.replace(',', ''))
        # Calculate the book-to-market ratio for each row
        book_value_data_BM['book_to_market_ratio'] = book_value_data_BM['book_value'] / book_value_data_BM['stock_price']
        return book_value_data_BM
```

Then we need to have time series market capitalization for the tickers *for constructing Size and Value Factors.* The data is then converted into a pandas DataFrame with the necessary columns, and the DataFrame is filtered based on the specified start and end dates. Finally, the function returns the filtered DataFrame which need for our time series data analyis.

Then I ranked the companies based on B/M and market capitalization to percentiles for each company based on their mean market capitalization weights [24] and B/M respective factors, and stored the result in a new DataFrame called "data". Specifically, it computes the rank percentile of each company's mean market capitalization weight in the distribution of all companies' mean market capitalization weights.

[24] $w_i = \frac{1}{N}$ value weight.

The code below defines a function named "get_market_cap" iterates through each tickers market cap and returns pandas data frame which is necessary for further filtrations. The data was in decedent order compared to B/M in required some cleanings and matching to have the data in order.

```
def get_market_cap(self):

    url = f'https://www.macrotrends.net/assets/php/market_cap.php?t={self.ticker}'

    response = requests.get(url)

    soup = BeautifulSoup(response.text, 'html.parser')

    script_tag = soup.find('script', text=re.compile('var chartData'))

    chart_data_string = re.search(r'var chartData = (.*);', script_tag.text).group(1)

    mktcap = eval(chart_data_string)

    df = pd.DataFrame(mktcap)

    df.rename(columns={'date':'Date', 'v1': 'Market Capitalization'}, inplace=True)

    df['Company'] = self.ticker

    df['Date'] = pd.to_datetime(df['Date'])

    df = df.loc[(df['Date'] >= self.start_date) & (df['Date'] <= self.end_date)]

    return df
```

Below function is called "download_stock_data" is designed to retrieve stock data for a list of given tickers within a specified time period. The function first creates two empty data frames, one for stock prices and another for nominal values. It then iterates through the given list of tickers, downloads the corresponding stock data using the "yf.download" function from Yahoo Finance, and extracts the adjusted close price and volume information. The function then calculates the nominal values by multiplying the adjusted close price with the corresponding volume. Finally, it concatenates the resulting data frames for prices and nominal values along the columns and returns them as output.

```
def download_stock_data(tickers, start_date, end_date):
    prices = pd.DataFrame()  # variable for stock prices
    nominals = pd.DataFrame()  # variable for nominals
    for ticker in tickers:
        stock_data = yf.download(ticker, start=start_date, end=end_date)
        stock_prices = stock_data[['Adj Close']]  # Adjusted Close Price
        stock_prices.columns = [ticker]  # Rename column as "Price of Company Ticker"
        # Copy the necessary columns to a new DataFrame to avoid SettingWithCopyWarning
        stock_nominals = stock_data[['Adj Close', 'Volume']].copy()
        # Calculate nominals without the prefix
        stock_nominals[ticker] = stock_nominals['Adj Close'] * stock_nominals['Volume']
        # Drop the original 'Adj Close' and 'Volume' columns
        stock_nominals.drop(columns=['Adj Close', 'Volume'], inplace=True)
        prices = pd.concat([prices, stock_prices], axis=1)
        nominals = pd.concat([nominals, stock_nominals], axis=1)
    return prices, nominals
```

The "download_stock_data" function can be used to retrieve stock data for a list of specified tickers within a specified time period. Then we use define class method to filter and clean the data, for further use "csv" file created and the below code shows and assures the time span are correctly applied.

```
prices=StockData.filter_dataframe_by_date_index(prices,start_date='2012-01-01',
end_date='2022-07-01')
nominals=StockData.filter_dataframe_by_date_index(nominals,start_date='2012-01-01',
end_date='2022-07-01')

prices.head()

#Output  of the data format is data fame in padans.
```

	PEAK	NTRS	BLK	CCL	... 250 tickers
2012-01-03	21.527918	31.058575	135.526749	25.530746	

This data can then be used to create *Momentum, Low Volatility, and Liquidity factors.* Each style factor is calculated based on third party data[25].

[25] ***For more details of data reaping and construction of factors please see the Class method.***

The research methodology involves analyzing the relationship between five factors (*Size, Value, Momentum, Low Volatility, and Sentiment)* and the outcome variables using both univariate and bivariate analysis. In the univariate analysis, portfolios are created for each factor using both equal-weighted and value-weighted [26]methods. The time-series averages of these portfolios are then cross-sectionally analyzed to identify statistically significant non-zero values in the difference portfolio. The first quintile represents the lowest values for each factor, while the fifth quintile represents the highest values. The values in each quintile are calculated using both equal-weighted and value-weighted methods.

For each traditional variable, 5 quintiles of portfolios are created and both equal-weighted (EW) and value-weighted portfolios are considered. Values within these portfolios are used to calculate their averages for each time period t and time-series of those averages are then cross-sectionally analysed. Statistically significant non-zero values of the outcome variables in the difference portfolio imply existence of a cross-sectional relationship between the explored variables.

In the bivariate analysis, only independent sorting procedure is utilised and only factorized portfolios are considered. Specifically, both of the sorting variables are categorised in three groups using percentiles 30 and 70 as the breakpoints. The following reporting standards are used:

- Size-Beta sort
- Size-Sentiment sort

The difference portfolios are again created, but this time labelled "3-1" since only three portfolios are considered in sorting per the sort variables. The time-series averages are reported in the same manner as it's done in the univariate analysis. The corresponding Newey-West t-statistics of the outcome variables and FFCPS alphas are reported.

The Fama & Macbeth regression is a statistical method that is commonly used in finance to control for the effects of multiple independent variables on the dependent variable. The method involves regressing the dependent variable (e.g. stock returns) on each independent variable (e.g. sentiment) independently, and then regressing the same dependent variable on all the independent variables together. The purpose of the FM regression is to determine the impact of each independent variable on the dependent variable after controlling for the effects of all other independent variables. This allows for a more accurate estimation of the effect of each variable on the dependent variable, as opposed to estimating the effects in isolation. The FM regression allows us to control for effects of all sort variables at once. Regressing the dependent variable on Sentiment independently is executed as well as regressing the stock returns on all the variables of interest all-together.

[26] Market capitalization (the number of shares outstanding is multiplied by the current market price per share) weighting also referred to as ***value weighting***, which is how the S&P 500 Index is calculated. Other widely used method is price-weighting method utilized by indices like the Dow Jones Index.

Empirical Analysis

This section describes how the proposed methodology was tested on actual data, which is crucial to confirm its validity of our hypothesis. The experiment included 250 sample of securities from stock exchanges in the S&P 500, and every step of the methodology was carefully examined. The results were then subjected to an out-of-sample validation process to ensure their accuracy. The empirical part is divided in three sections: the univariate analysis, the bivariate analysis and Fama&Macbeth regression and correlation seaborn map with the factorized sentiment index. The paper is mainly going to explore the relations of the following factors: *Size, Value, Momentum, Beta, and Sentiment*.

Developed a natural language processing algorithm to identify the content of companies' financial reports and use it to generate signals of firm performance. However, it is difficult to reflect various behavioral factors in a single variable hence utilized sentiment score composition with herding bias of BF.

- Company details: Information about the company, including its full name, domicile, unique identifier, securities identifiers, and others.
- Events: Information about the type of event detected in the news based on taxonomy.
- Scores: A set of numerical sentiment scores that represent different aspects of company-related news sentiment scores.

To conceptualize the herding behavior of the investors, I have used the below function to capture volume of the specific news targeted the crowd.

$$HEARDING\_BEHAVIOUR = \frac{1_D - average(1_D)}{\sigma_{daily}(1_D)}$$

Where:

1_D is daily count of the News.

σ_{daily} is the standard deviation of the daily volume.

$HEARDING\_BEHAVIOUR$ function that compares the daily volume with the average volume during the period specified.

In my more simplified model the function could potentially indicate herding behavior:

- *Increased Volume*: If the daily volume (Content_Volume_1D) is significantly higher than the average volume over a specified period, this could be indicative of herding behavior.
- *Increased Volatility:* If the standard deviation of daily volume is high, it may indicate increased volatility in trading volumes, which could also signal herding behavior.

Of coerce it is more simplified version to capture and open a quantified version of behavioral finance.

Below functions average of all of the daily averages over the timeframe specified. Nulls are excluded. Applies only to numerical fields. For example, if we wanted to calculate the daily average value for the Event Sentiment Score we would use the formula below:

$$\text{Average\_1D} = \sum_{i=1}^{n} \frac{News\ Sentiment\ Score}{n}, for\ i\ \epsilon\ U$$

Where U = {1, …, n} is the number of events for the company over the past 24 hours. Then we would apply the next formula below to calculate Strength.

$$Strength(LB) = \sum_{i=0}^{LB-1} \frac{w_{1,i*} w_{2,i*} Average\_1D}{\sum_{J=0}^{LB-1} If(Average\_1D = NULL)\ Then\ 0\ else\ w_{2,j}}$$

Where LB is the lookback window, $w_{1,}$ and w_2, are both exponential functions with different speed of decay. However, they take two different roles:

$w_{1,i}=2^{(-d_i/LB)}$ Here, d_i represents the number of days elapsed from the current day, ranging from 0 to LB-1, where LB denotes the lookback window. The weight factor $w_{1,i}$ acts as a degradation function, lessening the relevance of each event as a function of the time elapsed since the event occurred. Essentially, the function serves to diminish the influence of events that occurred in the more distant past, reflecting the notion that more recent events are usually of greater relevance.

$w_{2,i}=2^{(-10*d_i/LB)}$with d_i = 0 … LB-1 presenting a set of weights placing more importance on the most recent news thereby increasing the speed of decay for aged news. The variable d_i, ranging from 0 to LB-1, generates a set of weights via $w_{2,i}$ that assigns higher significance to the latest news, thus accelerating the decay rate for older news. This scheme emphasizes the importance of recent events while rapidly diminishing the impact of past ones. Furthermore, $w_{2,i}$ plays a crucial role in ensuring that the sentiment score is confined within the range of -1 to 1, thereby preserving the standard scoring framework for sentiment analysis.

In essence, the concurrent application of both $w_{1,}$ and $w_{2,}$ allows for an asymmetric weighting function that decreases the significance of news more swiftly for high news volume stocks (typically large cap) compared to low news volume stocks (typically small cap). Consider, for example, a company that experiences only a single event within a 91-day[27] window. In this scenario, only $w_{1,}$ is effectively operational, as $w_{2,}$ contributes only a single coefficient that is neutralized by the normalization component. In contrast, for a company that has multiple events within the same window, each event is assigned some weight, but not evenly. The weight assigned increases with the recency of the event, thereby enhancing the decay speed for older news.

[27] In many financial and business scenarios, key performance indicators and financial reports are analyzed on a quarterly basis.

Considering the above we will be more interested on size and sentiment factor in volatile environment.

The below formulas for calculating the expected return and variance of returns for a single security in two scenarios: when future returns are equally important, and when they're not. These concepts are essential in portfolio construction, where securities are combined in a portfolio.

Then for the return formula, let rij be the return of a security i during a time period j. The expected return $E(ri)$ of security i, assuming a series of M future time periods that are equally important, is defined as:

$$\mathrm{E}(r_i) = \frac{1}{M}\sum_{j=1}^{M} r_{ij}$$

If the future returns of a security cannot be considered equally important, the expected return $E(ri)$ can be calculated as:

$$\mathrm{E}(r_i) = \sum_{j=1}^{M} P_{ij}\, r_{ij}$$

whereP_{ij} is the possibility of return j for security i.

$$\sigma_i^2 = \frac{1}{M}\sum_{j=1}^{M} \left(r_{ij} - E(r_i)\right)^2$$

If the future returns of a security cannot be considered equally important, the variance of returns can be calculated as:

$$\sigma_i^2 = \sum_{j=1}^{M} P_{ij} \left(r_{ij} - E(r_i)\right)^2$$

where P_{ij} is the possibility of return j for security i.

In the context of our analysis, a portfolio refers to any combination of financial assets such as stocks, bonds, and cash, where each asset participates in the portfolio in some proportion that is determined by its value relative to the total value of the portfolio.

The portfolio standard deviation is defined as follows:

$$\sigma_{FP} = \sqrt{w_P^2\sigma_P^2 + (1-w_P)^2\sigma_F^2 + 2w_P(1-w_P)\rho_{FP}\sigma_P\sigma_F\backslash}$$

$$= w_P \sigma_P \Rightarrow w_P = \frac{\sigma_{FP}}{\sigma_P}$$

Beta:

Beta is used for the purpose of obtaining the standard beta from CAPM-based estimation, which measures the sensitivity to the market factor

$$\beta_i = \frac{Cov(R_{i,t}, R_{M,t})}{Var(R_{M,t})}$$

M … market related variable

i ∈ (1,250)

t is one time period

Size:

$$Size_{i,t} = \ln(MktCap_{i,t}),$$

MktCap … market capitalisation

i ∈ (1,250)

t is one time period

Momentum (MOM-proxy for mispricing patterns):

$$\text{Mom}(i,t) = \left[\prod m \in t-11; t-1(R(i,m)+1) - 1\right]$$

Where $\text{Mom}(i,t)$ $refers\ to\ the\ momentum\ factor\ of\ the\ i-th\ stock\ at\ time\ t$

The Mom equation is calculating the compounded return of stock i over the past 11 periods, which represents its momentum.

The security market line (SML) equation, which shows the relationship between the expected return of an asset and its beta coefficient:

$$E(r_i) = r_f + \beta_i(E(r_m) - r_f)$$

$E(r_i)$ is the expected return of asset i

r_f is the risk-free rate

β_i is the beta coefficient of asset i

$E(r_m)$ is the expected return of the market. Where $E(r_i)$and $E(r_m)$ are the expected returns on asset i and the market, respectively.

The formula for the Sharpe ratio, which measures the excess return of an asset over the risk-free rate per unit of its volatility:

$$\text{Sharpe ratio} = \frac{E(r_i) - r_f}{\sigma_i}$$

where σ_i is the standard deviation of the returns on asset i.

The factor model is a useful tool in finance and investment analysis that allows investors to understand the drivers of an investment's returns. By breaking down the expected return of an investment into the expected returns of underlying factors, the factor model provides a more nuanced and complete picture of the investment's risk-return profile. The only relevant factor in CAPM model is the return of the market, that is why CAPM is called a single-factor model. This information can then be used to make more informed investment decisions, such as allocating assets more efficiently or reducing exposure to certain risk factors.

If the general factor equation, is a regression model that predicts the expected return of a stock i, $E(r_i - r_f)$ based on the values of k factors *f*, the coefficients $\beta_{i,j}$ are the weights that are applied to each factor, and they are *estimated using a statistical technique called regression analysis*[28].

$$E(r_i - r_f) = \sum_{j=1}^{k} \beta_{i,j} f_j$$

Then we can modify it for factors:

$$E(r_i - r_f) = \sum_{j=1}^{k} \beta_{i,j} f_j = \beta_{i,1} f_1 + \beta_{i,2} f_2 + \cdots + \beta_{i,k} f_k$$

As a result, we will have the traditional factors final equation:

$$R_{stock} = \alpha + \beta_1 (R_{market} - R_f) + \beta_2 Size + \beta_3 Value + \beta_4 Mom + \beta_5 lowVol$$

For the sentiment factoring the equation will be as below:

$$R_{stock} = \alpha + \beta_1 (R_{market} - R_f) + \beta_2 Size + \beta_3 Value + \beta_4 Mom + \beta_5 Vol^{29} + \beta_6 Sentiment$$

The model is estimated using a linear regression. The dependent variable is the expected return of the stock, and the independent variables are the market return, the size of the stock, the value of the stock, the momentum of the stock, the volatility of the stock and Sentiment of the stock. The coefficient for sentiment (i.e. β6 in equation) would then represent the impact of sentiment on the returns of individual stocks, holding all other factors constant for the univariate analysis.

[28] The model is usually estimated using ordinary least squares regression, and the estimated coefficients represent the asset's sensitivities to the factors.

[29] Volatility has not included to the bivariate and univariate analysis.

Construction of the portfolios based on R_{stock} factors. The five variables of interest were explored: 5 portfolios were created for each of them, the average value in each time period (months) was calculated. The breakpoints for division into five portfolios are percentiles 20, 40, 60 and 80.

The **Table 2.01** below summarises the average values of the sorting variables in the quantile portfolios:

Table 2.01 Average Values of Sorting Variables for Portfolios

Average Values of Sorting Variables					
Sort Variable	***1***	***2***	***3***	***4***	***5***
ß	0.41859	0.779224	0.965875	1.14536	1.486498
Size	1.695191	2.54899	3.050766	3.683892	4.852949
Momentum	-0.06748	-0.014881	0.007301	0.029034	0.079276
Value	0.031029	0.194902	0.333764	0.538173	1.15322
Sentiment	-0.01367	0.064912	0.114148	0.166167	0.253793

(Source: Authors' analysis)

Beta is a measure of systematic risk, or market risk, and it tells you how much a particular security's price tends to move relative to movements in the overall market. The beta[30] is calculated based on monthly stock returns and market returns. It is computed as the ratio of the covariance between a given stock's returns and market returns to the variance of the market returns. The method generates beta values for each stock.

The **2.01 table** shows that the average beta of a portfolio increases as the quintile increases. This means that the portfolios in the highest quintile have the highest average betas. This suggests that the portfolios in the highest quintile are more volatile than the portfolios in the lower quintiles. Average values of size offer more to be thought about: the highest quantile reports only triple-the-value of the lowest quantile. That means, the market capitalisation of the 50 companies included in our top quantile is on average approximately 22 times larger than the average market capitalization of the companies included in the lowest quantile. However, we know the distribution is much more skewed when the whole market is considered. Nevertheless, the possible bias can be attributed to the sample selection since not including a very low number of the largest stocks can have a significant impact to the average values and distribution. The sentiment variable ranges from a high of 0.2533 to a low of -0.0136 between highest and lower quintiles. This implies that companies in the highest quintile have the highest level of sentiment score, while companies in the lower quintile have the lowest level of sentiment score. Sentiment can be thought of as a measure of the public's overall emotional state or sentiment towards the stocks. Therefore, the higher the sentiment, the more positive the market's overall emotional state towards the stocks in that quintile, and vice versa for lower sentiment levels.

[30] See the "class PortfolioAnalysis".

The averages of beta and momentum do not surprise since they are close global research values[31]. Apparently, these variables do not differ substantially from the whole market sample.

Here the goal of univariate analysis is to understand the distribution, central tendency, variability, and any outliers or skewness of the variable. This type of analysis is typically the first step in any data analysis and provides a preliminary understanding of the data, which is useful in the design of further, more complex analyses**.**

The stock returns were regressed on the five variables of interest – *Beta*, *Size, Value, Momentum, and Sentiment*. The risk-free rates of return were subtracted from the nominal returns of the stocks[32] in order to transform the plain data to excess returns. The one-month ahead excess returns are reported in percentages. Both equal-weighted and value-weighted portfolios were considered.

The following **Table 2.02** presents the results of equal-weighted univariate portfolio analyses of the relation between excess stock returns and the respective sort variables. The quintile breakpoints are percentiles 20, 40, 60 and 80 respectively. The excess returns are reported in p. m. percentages. Respective Newey- West t-statistics, adjusted using four lags, testing the null hypothesis that the difference portfolio excess return or FFCPS alpha is equal to zero, are reported in parentheses.

Table 2.02 Equal-Weighted Portfolios.

Equal-weighted portfolios: Excess returns							
Sort variable	**Portfolios: Quantile**					**Difference 5-1**	**FFCPS α**
	1	**2**	**3**	**4**	**5**		
ß	1.259228	1.382299	1.388657	1.237909	1.491521	0.242	-0.250
(t-stat)	*(3.588.)*	*(3.590.)*	*(2.513.)*	*(2.737.)*	*(2.499.)*	*(1.231.)*	*(-0.793)*
Size	1.537470	1.402812	1.359347	1.098122	0.902335	-1.404.	-0.910.
(t-stat)	*(4.54.)*	*(3.18.)*	*(2.30.)*	*(2.16.)*	*(2.26.)*	*(-5.661)*	*(-4.242*
Momentum	1.594236	1.303537	1.344931	1.182564	0.782160	-0.456.	-3.063.
(t-stat)	(2.479.)	(2.926.)	(3.576.)	(2.678.)	(2.623.)	(-0.933)	(-1.651)
Value	1.091627	1.045044	1.270932	1.475199	1.773834	0.682	-0.700
(t-stat)	(-0.532)	(2.429.)	(1.901.)	(0.073)	(2.678.)	(2.146)	(1.120)
Sentiment	1.001656	1.185785	1.230221	1.244341	1.331996	0.325	-1.200
(t-stat)	(2.091.)	(1.349.)	(0.889.)	(0.417.)	(-0.053)	(2.144)	(-2.050)

(Source: Authors' analysis)

The average values of the sort variables are statistically significant[33] in every quantile of Beta, Size and Momentum EW portfolios. However, it gets complicated with uncovering the pattern of the independent variables' effects on the excess return. In the first place, it confirms the notion that

[31] See https://quantpedia.com/strategies/value-and-momentum-factors-across-asset-classes/

[32] It refers to the percentage change in the value of the stock over a specific period.

[33] In a two-tailed test with a significance level of 0.05, a t-statistic with an absolute value of 2.0 or greater indicates statistical significance.

stock returns are influenced by a multitude of factors, not just a single variable. These factors contribute to the risk and return of stocks and are reflected in their excess returns. Although the excess returns appear to be driven by the independent variables themselves, the difference portfolios of beta and Momentum are lacking statistical significance to prove there is either a positive or a negative relation between the returns and the variable. This suggests that stocks that have recently performed well tend to have lower future returns.

Higher beta quintiles outperform lower beta portfolios, suggesting a positive relationship between beta and returns, as predicted by the CAPM. The FFCPS alpha for the 5-1 approach, on the other hand, is negative and statistically insignificant, implying that the extra return might be explained by FFCPS model variables.

Hence, the only statistically significant effect uncovered in the analysis is the effect of Size. Implying the smaller the firm, the larger the excess return, the result of our analysis is consistent with the general belief and the results of the analyses. In our sample, the Size effects appear to be both rather strongly statistically significant and also economically significant – the difference portfolio shows an implied 1.404% p. m. positive up-side for the low market capitalisation stocks. The Fama-French-Carhart-Pastor-Stambaugh alpha is also statistically and economically significant and implies negative 0.91% p. m. excess return non-determined by the independent variable, i.e., Size. High B/M ratio portfolios have higher excess returns than lower B/M ratios stocks which is 0.682% p.m. differs for 5-1. The value premium can be noticed clearly from the Table 2.02 that companies with high book-to-market ratios, also known as value stocks, outperform those with lower book-to-market values, known as growth stocks.

Looking at the table, the role of market sentiment in stock returns, while not evenly significant across all quantiles, however is noticeable and potentially important since the excess return of the portfolios are increasing towards positive sentiment. As it is noticeable there's a general increasing trend in excess returns from quantile 1 to quantile 5, implying that as sentiment increases, so do the returns. This finding indicates that sentiment is a useful factor in predicting excess returns. However, the trend of increasing returns with sentiment is visible, only the first quantile shows statistical significance which lowest sentiment quintile.

The t-statistic for the difference in excess returns between the top (5) and lowest (1) quantiles is 2.144. This value is greater than two, indicating that the difference in returns between these quantiles is statistically significant. This shows a considerable difference in sentiment-based returns, but given the unpredictability of the individual quantiles, this finding should be interpreted with caution.

The following table presents the results of value-weighted univariate portfolio analyses of the relation between excess stock returns and the respective sort variables. The quintile breakpoints are percentiles 20, 40, 60 and 80 respectively. The excess returns are reported in p. m. percentages. Respective Newey-West t-statistics, adjusted using four lags, testing the null hypothesis that the difference portfolio excess return or FFCPS alpha is equal to zero, are reported in parentheses.

Table 2.03 Value-Weighted Portfolios.

Value-weighted portfolios: Excess Returns							
Sort variable	**Portfolios: Quantile**					**Difference 5-1**	**FFCPS α**
	1	**2**	**3**	**4**	**5**		
ß	0.84	1.178	0.916	1.178	1.209	0.369	-0.368
(t-stat)	(-2.947)	(-3.538)	(-2.062)	(-2.468)	(-1.813)	(-0.557)	(-0.769)
Size	2.088	1.367	1.052	0.921	0.973	-1.115	-0.602
(t-stat)	(-4.267)	(-3.113)	(-2.25)	(-2.251)	(-2.595)	(-4.229)	(-.271)
Momentum	1.106	1.185	1.041	0.909	0.927	-0.179.	-1.802
(t-stat)	(-2.479)	(-2.926)	(-3.576)	(-2.678)	(-2.623)	(-0.933)	(-1.651)
Value	1.025	1.05	0.966	1.3507	1.045	0.02	-1.675
(t-stat)	(-1.545)	(-2.714)	(-2.94)	(-4.619)	(-3.117)	(-0.850)	(-1.675)
Sentiment	0.72	0.77	0.78	0.86	0.985	0.265	-1.45
(t-stat)	(-2.185)	(2.315)	(2.311)	(1.275)	(2.355)	(0.125)	(-1.307)

(Source: Authors' analysis)

In terms of statistical significance, very similar results are revealed by value-weighted analysis as well. Again, the returns appear to be driven by the independent variables, but the difference portfolios of beta and Momentum are lacking statistical significance to support the hypothesis of either a positive or a negative relation between the returns and the variable.

The only statistically and economically significant effect is the Size effect – implying there is a negative relation between the excess returns and market capitalisation of the firm. In comparison to the EW results, the average return in the lowest quintile is lower in the VW analysis. The high returns here and driven mainly by the smallest firms, hence by the stocks included in the lowest quintile. As these stocks are those with the lowest market capitalisation, value-weighting shifts the returns away from the lowest quintile partially. However, it is interesting that as the biggest stocks get more weight in VW analysis, their reported excess returns increase. This is unexpected since we thought the largest firms offered rather low returns mainly due to the frequency of trading with their stocks, hence restoring equilibrium trading prices quickly. Through objectively test this prediction and get a more concrete understanding of the inherently elusive notion of investor sentiment, we begin with a review of the peaks and declines in U.S. market sentiment from 2012 to the COVID pandemic, for example during the COVID-19 negative sentiment helped large cap firms to increase their excess returns on average. This summary is based on anecdotal evidence and hence, by its very nature, can only be a suggestive, categorization of sentimental variations.

Negative statistically significant FFCPS alpha implies there is a negative monthly excess return non-determined by Size of the stock. In other words, including other relevant independent variables to the regression specification may be beneficial in uncovering other relations. We suggest there is an up-ward bias in the Size's coefficient – including other variables may cause

lowering the coefficient of Size and pull the alpha towards zero level. When sentiment is low (not negative), As we can expect SMB equities have greater subsequent returns than big cap stocks, stocks with high return volatility than those with low return volatility, unsuccessful stocks over profitable firms, and nonpayers over dividend payers. However, cannot uncover those patterns with our study limits.

This **Table 2.03** is called a correlation matrix, and it provides information about the relationships between different investment factors.

Table 2.03 Correlation Matrix for Factors.

Correlation Matrix for Factors						
	Mkt Excess Return (Mkt-rf)	**Size**	**Value**	**Momentum**	**Beta**	**Sentiment**
Mkt Excess Return (Mkt-rf)	1	0.22	0.28	-0.25	-0.76	-0.06
Size	0.22	1	0.16	-0.12	-0.25	-0.03
Value	0.28	0.16	1	-0.14	-0.36	-0.07
Momentum	-0.25	-0.12	-0.14	1	0.43	0.04
Beta	-0.76	-0.25	-0.36	0.43	1	0.08
Sentiment	-0.06	-0.03	-0.07	0.04	0.08	1

Source: Authors' analysis).

I include sentiment as a factor in a multi-factor model and examine its correlation with other factors. **Table 2.03** displays the correlation coefficients between the factors and **Figure 1** shows that positive sentiment tent to depart from central tendency. Mkt-rf and Size (0.22), there is a low positive correlation between market returns (above the risk-free rate) and the size factor. This suggests that when the market performs better than the risk-free rate, the size factor also tends to increase, but not very strongly.

Mkt-rf and Beta variables (-0.76), there is a strong negative correlation between the market returns (over the risk-free rate) and the Beta variables. This suggests that when the market performs better than the risk-free rate, the Beta variables tends to decrease, and vice versa. Momentum and Beta (0.43), there is a moderate positive correlation between Momentum and Beta variables.

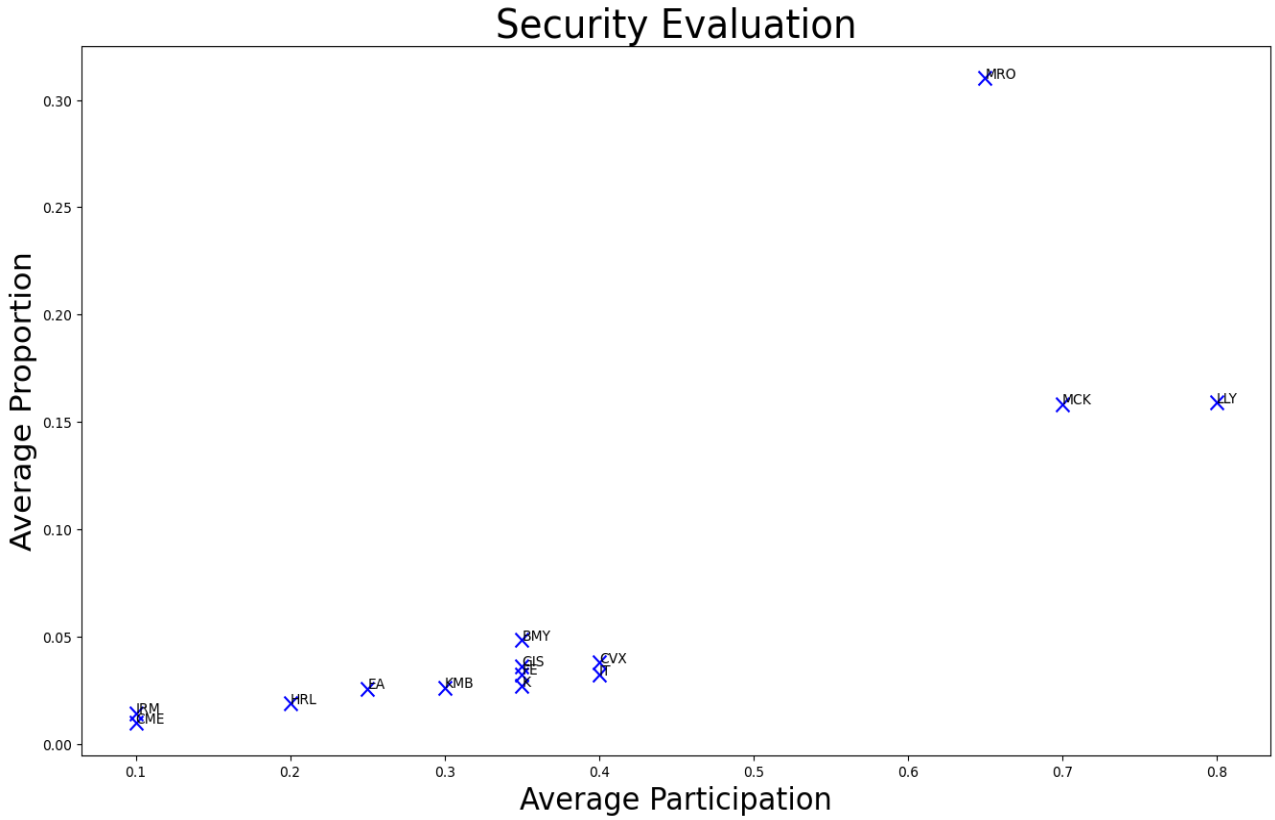
Market Excess Return and Value have a moderately favourable correlation (0.28). This means that when the market outperforms the risk-free rate, Value stocks tend to outperform as well.

This suggests that when one increases, the other also tends to increase, and vice versa. Value and Sentiment (-0.07), there is a very low negative correlation between the value and sentiment factors. This indicates that there's not a strong relationship in the movements between these two factors.

Our findings indicate that there is no significant linear relationship between sentiment variables and the other factors, suggesting that sentiment may be an independent factor in predicting stock

returns. These results can be new implications for investors and researchers interested in incorporating sentiment into factor models for asset pricing.

Figure 1. Security Evaluation (*Higher the average proportion is the most positive sentiment stock is).*



(Source: Authors' analysis)

We only utilised independent sorting in the analysis. The main variables of our interest– Beta, Size, Sentiment– were used for the following sorting combinations:

- Size – Beta
- Size – Sentiment

The bivariate sorting was executed in order to uncover any possible patterns of the excess returns' determination within the respective quantiles of the other sorting variable. Only value-weighted portfolios were considered. The breakpoints for division into three portfolios are percentiles 30 and 70.d

The following table presents the results of value-weighted bivariate portfolio analyses of the relation between excess stock returns, Size and Beta. The quintile breakpoints are percentiles 30 and 70 respectively. The excess returns are reported in p. m. percentages. Respective Newey-West t-statistics, adjusted using four lags, testing the null hypothesis that the difference portfolio excess return or FFCPS alpha is equal to zero, are reported in parentheses.

Table 2.04 Value-Weighted Portfolios.

Value-weighted portfolios: Bivariate independent sorting					
Sort variables	**Portfolios: Quantile**				**FFCPS α**
	ß1	**ß2**	**ß3**	**ß3-1**	
Size 1	1.645.	1.689.	2.187.	0.497.	0.036.
(t-stat)	*(5.615.)*	*(3.682.)*	*(3.194.)*	*(0.998.)*	*(0.110.)*
Size 2	0.924.	0.983.	1.156.	0.187.	-0.327.
(t-stat)	*(2.98.)*	*(2.19.)*	*(1.85.)*	*(0.44.)*	*(-1.059)*
Size 3	0.880.	1.002.	1.126.	0.202.	-0.021.
(t-stat)	(3.093.)	(2.429.)	(1.901.)	(0.417.)	(-0.053)
Size 3-1	0.8100.	0.7320.	1.1050.		
(t-stat)	(-4.950)	(-2.838)	(-3.587)		
FFCPS α	-0.518.	-0.371.	-0.078.		
(t-stat)	(-2.901)	(-0.860)	(-0.168)		

(Source: Authors' analysis)

Based on statistical significance of the 3-1 difference portfolio, there is a negative relation between the excess return and Size of the stock implied, controlled for beta. The Size effect is the strongest in the high beta quantile of stocks – hence for those stocks that are the most sensitive to the market movements. This is in line with theory – usually, *low market capitalisation stocks are believed to be rather less sensitive to the market* which in turn sentiment variables lose its impact for such companies. Therefore, the magnitude of the Size effect is driven by highly capitalised stocks' lower returns combined with their larger sensitivity to the market. Stocks with low market capitalisation show higher excess returns – an up-side 0.810 % p. m. *On the other hand, the empirical findings show that big enterprises would be less impacted by sentiment, and as a consequence, value weighting tends to conceal the significant trends*[34].

FFCPS alpha is statistically significant only in case of lowest beta quantile, which implies there is a negative 0.518% p. m. excess return that is not due to Size effect.

The difference portfolios of beta controlled for Size effect are not statistically significant at any quintile of Size. FFCPS alphas are not statistically significant at any quintile of Size, too.

The following **Table 2.05 presents** the results of value-weighted bivariate portfolio analyses of the relation between excess stock returns, Size and Sentiment. The quintile breakpoints are percentiles 30 and 70 respectively. The excess returns are reported in p. m. percentages. Respective Newey-West t-statistics, adjusted using four lags, testing the null hypothesis that the difference portfolio excess return or FFCPS alpha is equal to zero, are reported in parentheses.

[34] See https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2006.00885.x#fn1_1

Table 2.05 Value-Weighted Portfolios Sentiment.

Value-weighted portfolios: Bivariate independent sorting					
Sort variables	**Portfolios: quantile**				**FFCPS a**
	Sent 1	Sent 2	Sent 3	**Sent 3-1**	
Size 1	2.013.	1.864.	1.725.	-0.332.	-1.847.
(t-stat)	*(3.057.)*	*(4.617.)*	*(3.648.)*	*(-0.619)*	*(-1.172)*
Size 2	0.950.	1.154.	0.854.	-0.140.	-3.251.
(t-stat)	*(1.62.)*	*(2.84.)*	*(2.01.)*	*(-0.387)*	*(-2723)*
Size 3	0.893.	1.123.	0.832.	-0.106.	-1.600.
(t-stat)	(1.817.)	(3303.000.)	(1.810.)	(-0.245)	(-1.148)
Size 3-1	-1.1620.	-0.7840.	-0.9360.		
(t-stat)	(-3.836)	(-3.328)	(4.)		
FFCPS a	-3.406.	-2.467.	0.295.		
(t-stat)	(-3.373)	(-1.480)	(0.)		

Source: Authors' analysis)

Based on statistical significance of the 3-1 difference portfolio, there is a negative relation between the excess return and Size of the stock implied. The effect is the strongest in the low-Sentiment quantile for which also FFCPS alpha is statistically significant and economically significant – it implies -3.406% p. m. excess return explained by something else than Size effect controlled for Sentiment.

Moreover, the Size effect is present not only in the lowest Sentiment quantile but throughout the whole sample of analysed stocks. However, the pattern of it in unclear with regard to the magnitude of Sentiment– there is no apparent relation between the magnitude of Size effect and the magnitude of Sentiment.

The difference portfolios of Sentiment are revealed to be statistically non-significant, which means that controlled for Size, there is no relation between excess returns and Sentiment. A different situation emerges when it comes to FFCPS alpha – in the middle quantile of Size, it is statistically significant, which implies there is a negative excess return that is not due to the Sentiment effect when controlled for Size. The results on Sentiment's effect are relatively surprising – we expected to find that low market capitalisation stocks feature strong Sentiment effect and that is because those stocks are often less frequently traded than the largely capitalised stocks, which is directly linked to slower equilibrium reaching. Hence, the Sentiment effect may be able to persist for longer periods of time.

The following **Table 2.06** presents the results of Fama & Macbeth regression analysis of the relation between both the excess returns and Sentiment separately and excess return and all the variables all-together. The excess returns are reported in p. m. percentages. Respective t-statistics, testing the null hypothesis that the respective parameter is equal to zero, are reported in parentheses.

Table 2.06 Fama & Macbeth Regression.

Fama & Macbeth regression		
	1	**2**
Intercept	1.03	1.52
(t-stat)	*(2.27)*	*(4.16)*
Sentiment	-0.071	-0.056
(t-stat)	*(-1.264)*	*(-0.859)*
Beta		0.20
(t-stat)		*(0.42)*
Size		(0.29)
(t-stat)		*(-5.928)*
Adjusted R2	0.09	0.10
N	250	250

Source: Authors' analysis)

The findings reveal that the effect of Sentiment is statistically insignificant when used as the sole regressor in the analysis of returns. Notably, the regression indicates a statistically and economically significant excess return of 1.025% per month that cannot be attributed to the Sentiment factor alone.

Upon incorporating all the variables of interest into the regression model, the perceived insignificance of Sentiment as a determinant of stock returns behaviour is further underscored. The inclusion of additional variables into the regression model led to a decrease in the statistical significance of the Sentiment factor, suggesting that the effects previously captured by this variable were better attributed to other variables in this comprehensive analysis. The statistical insignificance of Sentiment and beta effects is thus highlighted.

The regression's intercept, which can be interpreted as the excess return that is not associated with any of the effects examined in the regression, has increased to a statistically significant 1.524% per month excess return. This regression model also provides further support for the statistical significance and negative correlation of the Size effect with future stock returns. The regression indicates that an increase of 1% in Size (market capitalisation) of a stock results in a decrease in its future return by 0.289%. Therefore, the findings suggest that as a stock's market capitalisation increases by 1%, its prospective stock return is expected to decrease by -0.289%.

Backtesting of Portfolios' Performance

The fabrication of factor models necessitates multiple data preparation stages coupled with meticulous modelling. This process can be demanding in terms of resources. These models are typically proprietary, constructed in alignment with a thesis's distinct comprehension of the basic attributes of securities and market dynamics. In many instances, risk evaluation is conducted by an isolated segment within the establishment, and traditional data-supplied factor models for risk approximation are utilized to scrutinize and communicate portfolio risk. It is noteworthy to observe that this framework can be effectively utilized in the backtesting of portfolios, providing critical historical perspective and performance insights for portfolios we created based on below criterias. Performance attribution predominantly adopts a retrospective approach—it helps for us in determining the contribution of specific factors to the portfolio's performance over a defined temporal span. These particular factors are those incorporated within the utilized factor model. This analysis provides valuable insights into the efficacy of investment strategies and decisions, enabling adjustments for future periods.

The **Table 2.07** outlines the summary of fundamental style and sentiment factors' empirical findings which utilized in the above empirical study. The below table presents four themes: Momentum, Size, Value, and Sentiment. Each theme has a corresponding raw factor and is equally weighted in the model. The table provides a brief explanation of how each factor is expected to influence stock performance.

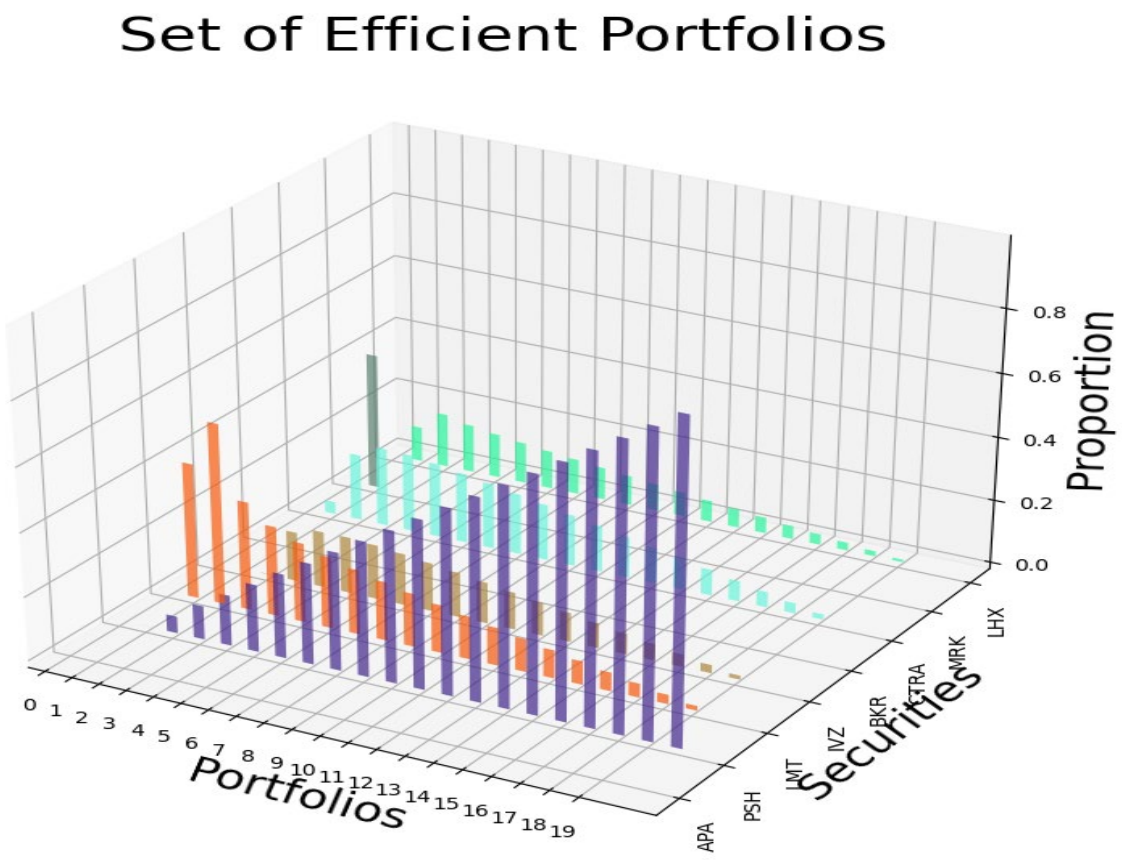
Table 2.07 Composition of Fundamental Style and Sentiment Factors.

Variables	Factors	Weight	Summary
Momentum	12-month return lagged 1 month	1/N	Stocks with high momentum are expected to perform better than those with low momentum and continue to display robust performance in the future. Momentum is frequently gauged through analysts' revisions and recent historical performance.
Size	$\ln(MktCap_{i,t})$	1/N	In the long term, it is anticipated that small-cap stocks will outperform large-cap stocks. This is assessed based on the company's market capitalization.
Value	Book to Market	1/N	Firms with high book-to-market ratios (commonly considered as 'value' firms) are expected to outperform those with low book-to-market ratios (typically categorized as 'growth' firms). Other common metrics used to evaluate value versus growth firms encompass earnings-to-price, free cash flow to price, and the dividend yield percentage.
Sentiment	Sentiment Score of Each Company	1/N	Companies scoring high on General Sentiment criteria on news should outperform companies with low Sentiment scores. Measured by using Sentiment scoring in BF criteria.

We will construct **table 2.07** Barra GM3 model[35] to test the performance of the portfolios.

Figure 2 is 3D display shows how the portolfios will be weighted and assests allocated to implment the above model. Please be noted it is only sample since there are 250 companies it is limited to fit those companies to show the reader how the allocattion was assigned.

Figure 2



The data was the same data was used for empirical part however the data cut off period reduced from 2022/07/01 to 2021/07/01 to capture the COVID-19 impacts. The edhec_risk_kit_206.py: [36] rich library was utilized to perform the backtesting of the portfolios. The function 'portfolio_stats' calculates various performance metrics for a given portfolio of assets, represented by their cumulative returns. The annualized return is calculated using the geometric mean of the portfolio's cumulative returns, and is expressed as a percentage. The annualized volatility is the standard deviation of the portfolio's returns, adjusted for the number of trading days in a year (252).

[35] See the https://www.msci.com/our-solutions/factor-investing

[36] See the https://github.com/suhasghorp/PortAnalyticsAdvanced/blob/master/edhec_risk_kit_206.py

Skewness and kurtosis are measures of the shape of the distribution of portfolio returns. Skewness indicates whether the distribution is symmetrical (skewness=0), positively skewed (skewness>0) or negatively skewed (skewness<0). Kurtosis measures the degree of peakedness or flatness of the distribution, with higher values indicating more extreme values (heavy tails).

The Cornish-Fisher Value-at-Risk (VaR) at the 5% level is calculated using the normal distribution, based on the portfolio's mean and standard deviation. Historic Conditional VaR (CVaR) at the 5% level is calculated as the average return of the worst-performing 5% of portfolio returns.

The Sharpe Ratio measures the portfolio's risk-adjusted return, and is calculated as the excess return over the risk-free rate (assumed to be 2.5%[37]) divided by the portfolio's annualized volatility.

Finally, the maximum drawdown is the largest percentage decline from the portfolio's previous peak, and is an indicator of downside risk for the investors. The below given print screen represents the annualized returns, annualized volatility, skewness, kurtosis, Cornish-Fisher Value-at-Risk (VaR) at the 5% confidence level, historic Conditional Value-at-Risk (CVaR) at the 5% confidence level, Sharpe Ratio, and maximum drawdown of first ten different tickers. The returns of the stocks range from negative 0.134 to 0.30 with a mean of 0.116 and a standard deviation of 0.172. The volatility of the stocks ranges from 0.146 to 0.81 with a mean of 0.34 and a standard deviation of 0.20. The skewness and kurtosis of the data indicate the degree of asymmetry and peakedness of the return's distribution, respectively.

```
In [23]: erk.summary_stats(stock_returns_monthly)
```

Out[23]:

	Annualized Return	Annualized Vol	Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Historic CVaR (5%)	Sharpe Ratio	Max Drawdown
LRCX	0.301130	0.314214	0.131538	3.009922	0.119042	0.146106	0.839587	-0.439111
SYY	0.126085	0.223491	-0.242265	10.575447	0.088194	0.136868	0.418161	-0.463744
VTRS	-0.006225	0.325595	-0.528245	4.296003	0.161056	0.214670	-0.108714	-0.799807
CRL	0.201686	0.273070	-0.077175	3.256972	0.111983	0.149384	0.611643	-0.352216
UAL	0.100029	0.385149	-0.507793	5.424135	0.177721	0.237531	0.176381	-0.710031
LDOS	0.142250	0.271778	0.261264	4.539085	0.106074	0.153666	0.401666	-0.430579
JPM	0.148212	0.253684	-0.499045	4.174738	0.113980	0.161528	0.453220	-0.350015
HSY	0.158248	0.172483	0.478475	4.699053	0.059411	0.087041	0.723533	-0.198338
AWK	0.225103	0.146104	0.103266	3.924063	0.049156	0.073510	1.299731	-0.135202

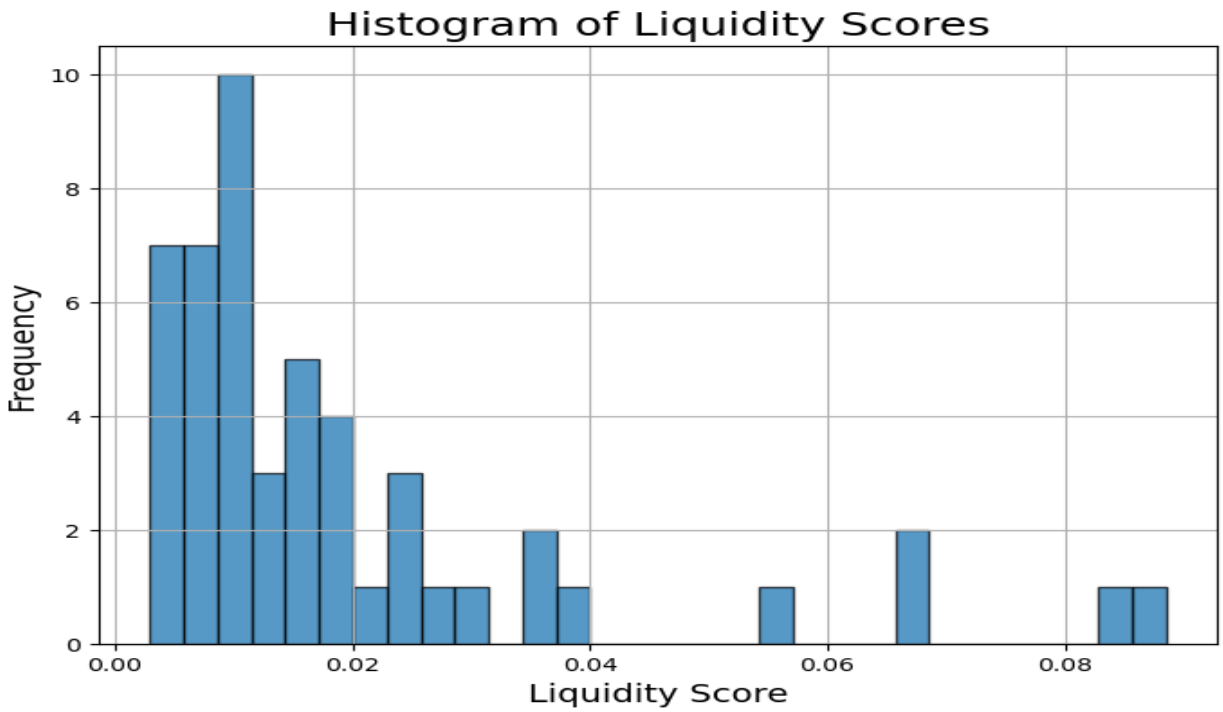
The output of the function is a panda's series containing above performance metrics. The function(***edhec_risk_kit_206.py)*** "backtest_ws" is designed to perform a backtest of a given weighting scheme on a set of asset returns. It takes several parameters including the asset returns (r), an estimation window length (estimation window), and a weighting scheme (**weighting Figure 2**) that is specified as a function of the asset returns and any additional keyword arguments. The function first divides the asset returns into a series of windows of length "estimation window", and then uses the weighting scheme to determine the weights to assign to each asset in each window.

[37] See https://ycharts.com/indicators/3_month_t_bill

These weights are then used to calculate the returns for each window, which are combined to generate the overall returns for a portfolio. The function returns a pandas DataFrame containing the portfolio returns for each period in the estimation window which is to fine for comparing the cumulative return of portfolios.

The code creates a histogram of liquidity scores using Matplotlib in Python, which shown **in Figure 3** where the liquidity scores are plotted on the x-axis and their frequency on the y-axis. The tick labels on the x-axis display the ticker symbols, and this liqudity scores then are utilized to contrcut the liqudty weigthed portolios and calcaulate its cumulative returns as well.

Figure 3

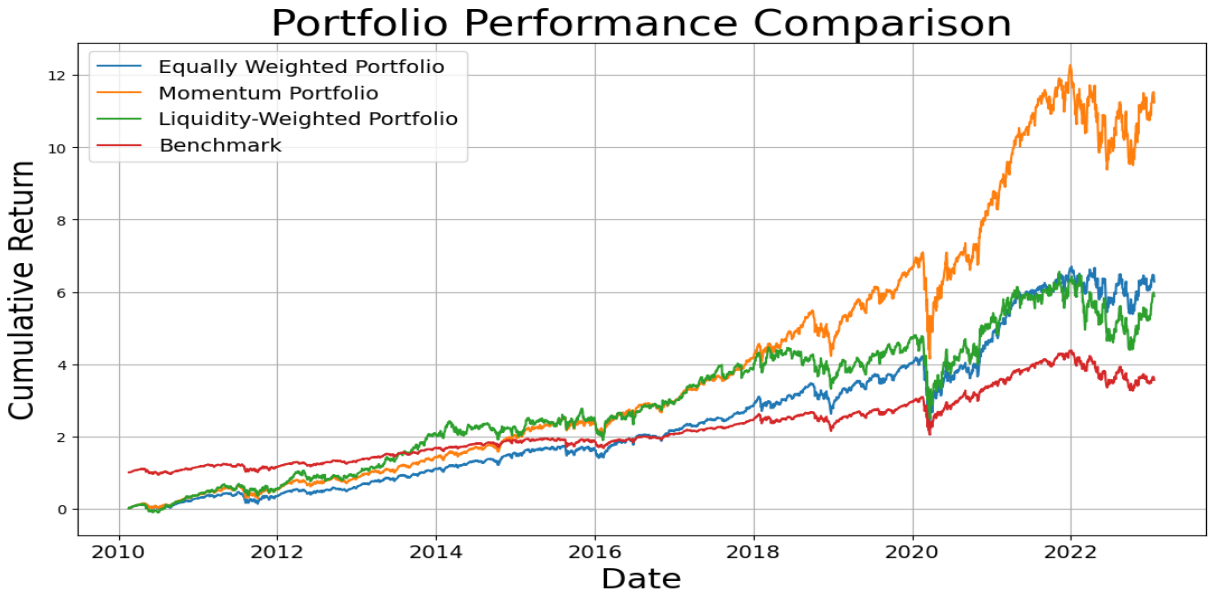


The cumlative return comparision of factor based portolios displays the performance metric of: Equally Weighted, Momentum, and Liquidity-Weighted portolios with comparisosn to Benchmark(S&P 500 index). The annualized return for Momentum is the highest at 0.21, followed by Equally Weighted at 0.166 and Liquidity-Weighted at 0.161. However, Liquidity-Weighted has the highest annualized volatility at 0.232, followed by Momentum at 0.190 and Equally Weighted at 0.184. Skewness, which measures the degree of asymmetry in the distribution of returns, is negative for all three strategies, with Momentum exhibiting the lowest skewness at -0.2940. Kurtosis, which measures the degree of peakedness in the distribution of returns, is highest for Equally Weighted at 12.23, followed by Momentum at 10.78 and Liquidity-Weighted at 6.83. The Cornish-Fisher VaR (Value at Risk) at a 5% significance level is lowest for Liquidity-Weighted at -0.167, followed by Equally Weighted at -0.134 and Momentum at -0.231. The Historic CVaR (Conditional Value at Risk) at a 5% significance level is lowest for Equally Weighted at -0.028, followed by Momentum at -0.028 and Liquidity-Weighted at -0.034. The Sharpe Ratio, is highest for Momentum at 1.01, followed by Equally Weighted at 0.746 and Liquidity-Weighted at 0.607.

Finally, the Max Drawdown, is highest for Liquidity-Weighted at -0.393, followed by Equally Weighted at -0.381 and Momentum at -0.364. Overall, the ourcome suggests that Momentum has the highest risk-adjusted return, followed by Equally Weighted and Liquidity-Weighted.

Table 2.08 Traditional Factors (Cumulative Returns).

	Equally Weighted	Momentum -Weighted	Liquidity-Weighted
Annualized Return	0.166254	0.214126	0.161192
Annualized Volatility	0.184841	0.190865	0.23251
Skewness	-0.396953	-0.294031	-0.259397
Kurtosis	12.239027	10.787538	6.837499
Cornish-Fisher VaR (5%)	-0.13455	-23%	-0.161057
Historic CVaR (5%)	-0.028093	-3%	-0.034859
Sharpe Ratio	0.791246	1.017087	0.60725
Max Drawdown	-0.382461	-0.36409	-0.398853



For the summary of above **Table 2.07**, I would say momentum provides the highest cumulative return, but for this reward also it has the highest risk, as illustrated by its volatility and maximum drawdown. Liquidity-Weighted portfolio delivers lower returns and more risk. However, the Equally-Weighted approach has a more balanced risk/reward profile, but its maximum drawdown remains high, implying that there may be periods of significant losses. As a result, investors must consider these aspects against their risk tolerance and investing objectives.

The process of developing a portfolio based on S&P 500 constituents entails selecting U.S. corporations that meet particular market capitalization criteria such as considering survivor bias.

I have analyzed for further performance analysis of three more different portfolio strategies: GMV, Low Volatility, and Market Cap. The annualized return for each portfolio is given, as well as the annualized volatility. As a result, the Low Volatility portfolio had the highest Sharpe ratio of 0.811, indicating the best risk-adjusted performance, followed by the Market Cap portfolio with a Sharpe ratio of 0.823. The GMV portfolio had the lowest Sharpe ratio of 0.749. In terms of annualized

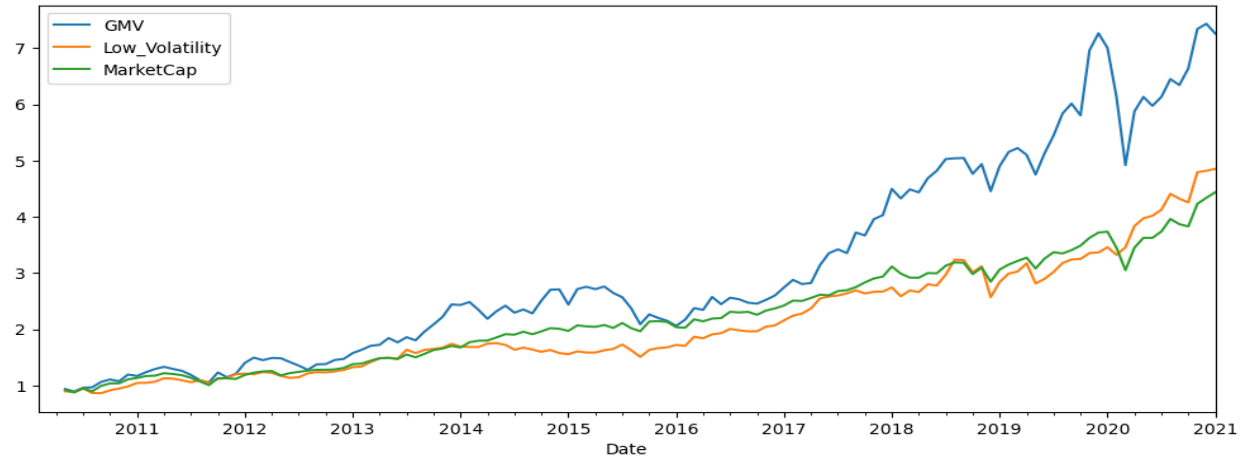
returns, Low Volatility had the highest return of 0.158, followed by Market Cap with a return of 0.149, and GMV with the lowest return of 0.203.

In terms of risk, Low Volatility had the lowest annualized volatility of 0.154, followed by Market Cap with a volatility of 0.140, and GMV with the highest volatility of 0.224. Low Volatility also had the lowest maximum drawdown of -0.207, followed by Market Cap with a maximum drawdown of -0.183, and GMV with the highest maximum drawdown of -0.323.

These outcome of that among the considered portfolio construction methods, the Low Volatility strategy had the highest Sharpe ratio and the lowest risk, indicating a better risk-adjusted performance and lower downside risk compared to the other strategies. The Market Cap strategy had a higher Sharpe ratio and annualized return than the GMV strategy, but also had higher volatility and drawdown. The GMV strategy had the lowest Sharpe ratio and highest volatility and drawdown among the three strategies.

Table 2.09 Traditional Factors (Cumulative Returns).

	Annualized Return	Annualized Vol	Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Historic CVaR (5%)	Sharpe Ratio	Max Drawdown
GMV	0.2025	0.22411	0.03192	3.58636	0.08712	0.11346	0.74897	-0.32304
Low_Volatility	0.1583	0.15387	-0.68235	5.45451	0.06547	0.09709	0.81141	-0.20674
MarketCap	0.14867	0.14027	-0.1009	4.01355	0.05425	0.07725	0.82332	-0.1834



For the **table 2.09**, the Sentiment score weighted portfolio has the highest annualized return and Sharpe ratio, with a moderate level of volatility and downside risk. This suggests that it may be a rational choice for investors seeking higher returns with reasonable risk. Looking at the annualized return, the Sentiment factor seems to have the highest return of 0.245, followed by LowSize with a return of 0.205, Value with a return of 0.204, and EW with a return of 0.19 This suggests that investors who prioritize high returns may prefer a portfolio that includes the Sentiment or LowSize factor.

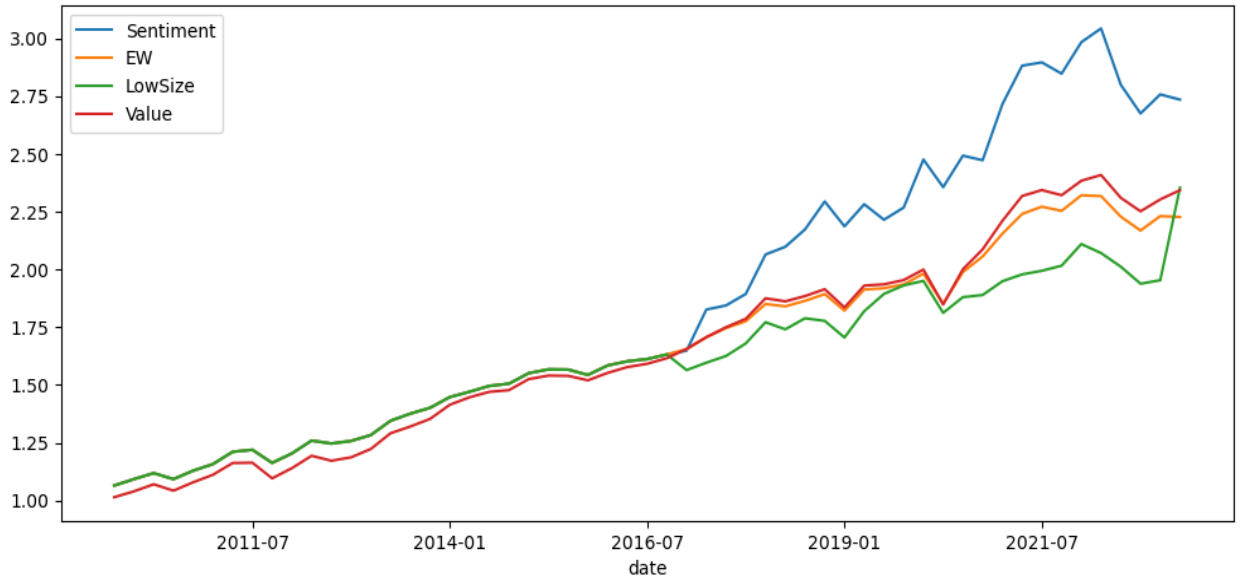
On the other hand, we also need to consider the risk of each factor. Annualized volatility, is highest for LowSize at 0.132, followed by Sentiment at 0.1280, Value at 0.100, and EW at 0.092. Sentiment has the highest max drawdown at -0.120915, followed by LowSize at -0.081, Value at -0.075, and EW at -0.065. *This implies that investors who prioritize avoiding large losses may prefer to rebalance their portfolios continually depending on sentiment of the financial market towards the specific stocks or portfolios*.

The Sharpe ratio, is highest for EW at 1.692, followed by Sentiment at 1.637, Value at 1.686, and LowSize at 1.288.

Looking at the Cornish-Fisher VaR and historic CVaR, which measure the expected worst-case loss of a portfolio, we see that LowSize has the lowest values at 0.016 and 0.052, respectively. This suggests that investors who prioritize downside protection may prefer a portfolio that includes the LowSize factor. Sentiment or LowSize factor. The Sentiment strategy had the highest annualized return and Sharpe Ratio, indicating superior performance.

Table 3.00 Traditional Factors and Sentiment Factor (Cumulative Returns).

	Annualized Return	Annualized Vol	Skewness	Kurtosis	Cornish-Fisher VaR (5%)	Historic CVaR (5%)	Sharpe Ratio	Max Drawdown
Sentiment	0.245435	0.128097	-0.05294	3.564139	0.041281	0.058443	1.637036	-0.12092
EW	0.166254	0.184841	-0.66739	12.2390	0.032847	-0.028093	0.791246	-0.382461
LowSize	0.205373	0.132522	1.81931	12.35215	0.016869	0.05294	1.288034	-0.08137
Value	0.204125	0.100485	-0.84388	4.365374	0.036982	0.058303	1.686687	-0.07562



Conclusion

The central idea is that stocks that are difficult to value using fundamental analysis are also difficult to arbitrage. In this study, we investigate the potential benefits of incorporating news sentiment variables from ML Analytics into traditional style factor indices for the constituents of cap weighted S&P 500 index. Sentiment Analysis, as we know, employs text mining techniques to determine the views and sentiment expressed in news and postings, classify their polarity, and/or provide a sentiment score. On the other hand, Behavioral Finance is examining more subjective and qualitative part of the finance, and it is hard to quantify and measure its well-documented behavioral bias concepts' impact on time varying stock returns. However, we used sentiment variables to shape and fit the BF biases into our model and to measure its impact. And our finding is mostly related to randomly sampled 250 stocks out of 500, S&P 500 components. We did not divide the companies into its sectors. For instance, communication sector of the S&P 500 may be more influenced by the sentiment of the public. However, we only focused on factor variables which rooted back by Fama and French, as well as Stephen Ross.

We build a total of four traditional style factor, as well as an alternative sentiment factor using news text analysis, empowered by AI that is why we called sentiment proxy for AI.

For our study limit and based on our empirical findings we developed below Hypothesis:

1. Hypothesis: Artificial intelligence-based sentiment index is a rational tool for investors.

2. Hypothesis: While markets interact with investors, an inherent mapping exists between investor sentiment and market conditions that reveals future market trajectories.

3. Hypothesis: Investment decision is influenced by psychological and emotional factors.

To check the ***Hypothesis.1.*** we need to dive deep to quintile univariate analysis, in our study we found an asymmetrical relationship when the sentiment variables were split into negative to positive sentiment scores according to quintiles. The positive sentiment factors have a positive effect on excess market return, but the intensity of negative sentiment has a negative effect on negative returns. These findings show that when investors are more confident about the market providing excess returns, their extreme optimism leads to speculative actions that urge them to follow the crowd. The study also discovered the persistence of market volatility and sentiment, demonstrating the contemporaneous impact on sentiment and excess market returns. We also witnessed that size effect can be noticed in every part of our study we strongly believe it is because of solid public view about those companies thus they have established reputation and rock-solid balance sheet which do not offer higher premiums for the investors in turn of risk.

On the other hand, the correlation map of the variables of interest revealed that when the market outperforms the risk-free rate, Value stocks tend to outperform as well. This is because, during a downturn, investors are inclined toward growth stocks, which might lead value equities to become undervalued. We believe it's because of a risk appetite of the investor, as macroeconomic indicators change investors risk preferences change as well. In the downturn, and in the volatile uncertain environment investors look for certainty and less risky, too big to fail companies to invest in. This is same as Beta that higher market returns above the risk-free rate coincide with a lower Beta which means market is doing better in low beta conditions and beta can be informative tool for the extreme risk averse investors. Surprisingly Sentiment variables seem to have very low correlation with other factors. This could imply that sentiment is relatively independent from the other factors and might be driven by different market conditions or investor behaviors. Nevertheless, it can be meant for the investors that is more informative and has low correlation with other factors as well has stable persistent return. We believe the Sentiment can be utilized by investors as well as companies to detect public view towards to them and control the risk associates with certain public behavioral biases. Thus, it can be rational defensive tool for investors.

To elaborate our findings with the ***Hypothesis.2.*** we need to investigate the bivariate analysis and its outcomes alongside correlation matrix of variables. Controlling for Beta, there appears to be a significant negative relationship between the size and its excess return. This association appears to be especially strong for equities with high Beta values, indicating that larger companies with high sensitivity to market movements likely to have lower excess returns. For Size and Sentiment similarly, a significant negative relationship is found between the Size of a stock and its excess return when controlled for Sentiment. Notably, this effect is statistically significant across all sentiment quintiles, but there is no clear link between the magnitude of the size effect and the intensity of the size effect. We also noticed that Sentiment, when controlled for Size, doesn't seem to have a statistically significant effect on excess returns. This was unexpected for us as the hypothesis was that smaller, less frequently traded stocks would have a more pronounced sentiment effect. Similarly, Beta didn't show a statistically significant relationship with excess returns at any quintile of Size. The FFCPS alpha is only statistically significant in the case of the lowest beta quintile, implying there is a negative 0.518% p.m. excess return that is not due to the Size effect. This implies that there may be other factors influencing these returns. When it comes to Fama & Macbeth Regression, sentiment was used as the sole regressor in the analysis of returns, it was not found to be statistically significant. The inclusion of additional variables (Size and Beta) into the regression model further decreased the significance of the Sentiment factor. Size had a significant negative impact on the returns. The insignificance of the Sentiment and Beta in the multivariate regression analysis was also noted. Thus, we failed to prove the ***Hypothesis.2.*** that future market conditions cannot be revealed only with sentiment variables even though sentiment variables created by people's interaction there is some other factors which should be considered.

The conclusion from the bivariate analysis is that the Size effect, as measured by market capitalization, has a significant and negative impact on future stock returns. The results also indicate that there is no significant relationship between future stock returns and beta or sentiment when controlled for Size. The negative FFCPS alphas suggest that the low capitalization stocks outperform high capitalization stocks.

Moreover, news associated with stocks, such as news stories or analyst reports, might impact the stock returns, meaning that the stocks continue to move in the direction of the initial news for a period of time after the news is released, however quick shocks are easily recovered and do not affect in the long-term return of the portfolios. The decay period of the news is 91-days since after this period news become worthless for investors.

In the Barra GM3 model we compared the following strategies: GMV, Momentum, Equally Weighted, and Liquidity-Weighted, Low Volatility, Market Cap, Value, LowSize and Sentiment. With the highest annualized return and Sharpe Ratio, the Momentum strategy outperforms the others in terms of risk-adjusted performance. However, it also exhibits high volatility and maximum drawdown, indicating potential risk. Furthermore, the GMV, Low Volatility, and Market Cap methods. Particularly, the Low Volatility strategy outperforms in terms of risk-adjusted performance, with the highest Sharpe Ratio and the lowest risk indicators like annualized volatility and maximum drawdown. The GMV approach appears to have the highest return, but it also has the largest volatility and maximum drawdown, making it less appealing to risk-averse investors.

The Sentiment strategy likely to be the best performing with the highest annualized return and a competitive Sharpe Ratio. However, it carries the highest max drawdown, suggesting potentially significant losses in some periods. Wwhen we add news sentiment to a traditional multifactor model with a monthly rebalancing frequency, we observe significant improvements, based on the **Table 2.09** to check the ***Hypothesis.3.*** we can say that sentiment has 24% annualized cumulative return and it outperformed for 400 b.p. other factors. However, it worth to mention that for sentiment score weighted portfolios need to be continually rebalanced most precisely monthly and quarterly depending market conditions and locations. If sentiment is added to these factors, the strategy shows further gains in return, without increasing risk.

Gains followed by Value and Momentum incorporating sentiment variables did not significantly alter exposure to the targeted factor, but resulted it is informative.

The study was able to arrive to several conclusion which are following and are divided into parts that respect the previous structure of the study. General remarks include the following:

- The regressions specifications utilised in the study are not complex and statistically powerful enough to uncover the true relations governing the behaviour of excess returns at the financial markets.
- Size effect appears to be rather strong and persistent both in time and among other independently assessed effects of other variables. The relation between the market capitalisation of the stocks and future excess returns in negative and sentiment effect is minor for large cap stocks.
- Fama-Macbeth regression supports the previously solicited statistical evidence for Size effect having negative effect on future returns of the stocks.
- Sentiment variables are forward looking and is based on expectations of a company's future performance.
- Sentiment variables can be useful tool for the investors to track the behavioural risks.
- AI based tools can be customized based on above models.

Table 3.01 Summary of the Sentiment Variables.

Sentiment Factor Contribution	
Objective	**Description**
Addressing values-based investment constraints	Screening or avoiding certain securities based on ethical, social, or religious considerations.
Mitigating long-term systemic risks	Addressing risks that may have spillover effects on the economy or society as a whole.
Reducing systematic risks caused by changes in markets	Mitigating risks that may arise from changes in the market environment or the economy.
Identifying stock-specific opportunities and risks	Analyzing individual securities to identify potential opportunities and risks.
Investing with impact of AI based sentiment signals that outperform the S&P 500 benchmark by significantly	Investing in companies or assets that have a positive social or environmental impact.

Our main findings suggest that news sentiment can be a robust source of medium-term return premia for the S&P 500. In fact, the standalone sentiment index that we create is uncorrelated with traditional risk factors and yields more than 400 basis points in annualized returns, even for the monthly rebalancing portfolios.

In general, our results suggest that incorporating news sentiment data from sentiment variables can enhance the performance of traditional style factor indices and lead to better portfolio outcomes for investors in the S&P500. The relationship between sentiments and financial market returns tends to concentrate on whether sentiment data can be used to predict stock market movements rather than determining a causal link between the two. The difficulties in making inferences about individual investor behaviour based on aggregate data and the possibility of spurious or vanishing causal relationships found are the reasons for this focus. Moreover, sentiment data may only represent an interpretation of market fundamentals rather than having a causal impact on market movements. Despite these limitations, sentiment data can still be beneficial to investors as it reflects the average interpretation of market fundamentals by investors and may enhance trading strategies.

The focus of the work being discussed is on understanding the relevance of behavioural finance (BF) for individual investors. It investigates the aspects of BF that challenge the assumption of rational market behaviour and highlights cognitive biases. It also suggests ways that investors can use BF in their own investments. In summary, this research emphasizes the significance of considering BF when making sentiment analysis and investment decisions.

***Our model can be used in the future**:*

1) Advent of new AI cost effective solution to tune ML models such as PaLM 2 (Google's next-generation LLM), Model Garden and other currently 60+ models to uncover the hidden patterns between stock market and behavioural finance.
2) The Sentiment Model can be easily acquired over Refinitiv Eikon's Thomson Reuters MarketPsych Indice which is costly however user friendly and easily be manipulated for users' input variables.
3) Investors and academic background individuals can develop the more advanced and costumed model which can yield significant return and outcome since we know from our empirical findings that sentiment variables are significant variables in 90 days window (decay period).
4) Investors can use company sentiment scores in order to build classic factor strategies. Incorporating sentiment variables resulted in higher returns while decreasing risk.

Appendix I:
Filters for Sentiment Event Detection(This information is confidential)

Field Name	Explanation
Analyst_Ratings_Sentiment	The sentiment of an analyst rating in the document (positive, negative, or neutral)
Category	The category associated with the document.
Commentary_Sentiment	The sentiment of the document in relation to commentary.
Composite_Sentiment_Score	The composite sentiment score of the document.
Confidence_Level_for_Entity_Sentiment	The confidence level of the sentiment score assigned to the detected entity.
Confidence_Level_for_Entity_Text_Sentiment	The confidence level of the sentiment score assigned to the text associated with the detected entity.
Corporate_Actions_Sentiment	The sentiment of the document in relation to corporate actions.
Country_Code	The country code associated with the entity.
Document_ID	The unique identifier assigned to each document by random number generater
Document_Sentiment	The overall sentiment of the document.
Document_Sentiment_Confidence	The confidence level associated with the document sentiment score.
Document_Type	The type of document, such as news article, press release, or financial statement.
Earnings_Release_Sentiment	The sentiment of an earnings release event in the document (positive, negative, or neutral)
Earnings_Tone_Sentiment	The sentiment of the document in relation to earnings.
Entity_Detection_Distance	The distance between the detected entity and the entity in the original text.
Entity_Detection_Type	The type of detection method used to identify the entity.
Entity_Hierarchy_Level	The level of the entity in the entity hierarchy.
Entity_ID	The unique identifier assigned to each entity in the document by random number generater
Entity_Name	The name of the entity detected by code.
Entity_Relevance	The relevance score of the detected entity.
Entity_Sentiment	The sentiment score of the detected entity.
Entity_Text_Sentiment	The sentiment score of the text associated with the detected entity.
Entity_Type	The type of entity detected by codes. Examples include company, person, organization, etc.
Evaluation_Method	The evaluation method used to determine the sentiment score assigned to the document.

Event_Detected_Entity_ID	The unique ID of the entity being discussed in the document where the event was detected.
Event_Detected_Entity_Name	The name of the entity being discussed in the document where the event was detected.
Event_Detection_Distance	The distance (in number of words) between the event and its detection within the document.
Event_End_Date_UTC	The UTC date and time when the event being described in the document ended.
Event_Risk	The level of risk associated with the event being described in the document.
Event_Sentiment	The sentiment of the event being described in the document.
Event_Similarity_Days	The number of days used to calculate event similarity.
Event_Similarity_Key	A key used to identify similar events across different documents.
Event_Start_Date_UTC	The start date of the event associated with the document
Event_Text	The text of the event being described in the document.
Fact_Level	The fact level associated with the document.
Group	The group associated with the document.
Interest_Rate_Sentiment	The sentiment of the document in relation to interest rates.
Maturity	The maturity level of the document.
Mergers_Acquisitions_Sentiment	The sentiment of the document in relation to mergers and acquisitions.
Original_Language	The original language of the document.
Parent_Entity_ID	The unique identifier assigned to the parent entity of the detected entity.
Parent_Entity_Name	The name of the parent entity of the detected entity.
Position_Name	The name of the position associated with the document.
Provider_Document_Chain_ID	The unique identifier of the document chain assigned by the provider
Provider_Document_ID	The unique identifier of the document assigned by the provider
Provider_ID	The unique identifier of the provider that delivered the document
Related_Entity_ID	The unique ID of the related entity being discussed in the document.
Related_Entity_Name	The name of the related entity being discussed in the document.
Relationship	The type of relationship between the entity mentioned in the document and the related entity.
Reporting_End_Date_UTC	The UTC date and time when the reporting period ended.
Reporting_Period	The time period covered by the document's content.
Reporting_Start_Date_UTC	The UTC date and time when the reporting period started.
Reporting_Type	The reporting type of the document.
Role	The role of the document.
Territory_ID	The unique identifier assigned to the territory

Segment_Name	The name of the segment associated with the document.
Sentiment_Impact_Projection	The projected impact of the sentiment of the document on related entities.
Source_ID	The unique identifier of the news source that published the document
Source_Name	The name of the news source that published the document
Source_Rank	The rank of the news source based on factors such as reach and credibility
Stock_Tone_Sentimen	The sentiment of the document in relation to stocks.
Sub_Type	The sub-type of the document.
Sustainability_Sentiment	The sentiment of the document in relation to the topic of sustainability.
Territory_Name	The name of the territory associated with the document.
Title	The title of the document.
Title_Similarity_Days	The number of days used to calculate title similarity.
Title_Similarity_Key	A key used to identify similar titles across different documents.
Topic	The topic associated with the document.
Type	The type of document.
Unit	The unit associated with the document.
UTC_Timestamp	The timestamp of the document in UTC format.
Word_Count	The number of words in the document

***Formulas can be utilized for above filter*:**

cardinality	Counts the distinct values for a field for each day. Nulls are excluded. Cardinality does not support a window size.	Numerical fields
count	Counts the values for a field for each day. Nulls are excluded.	Numerical fields
daily_avg	Calculates the average of all daily averages over the specified timeframe. Nulls are excluded.	Numerical fields
dense_rank	Consecutively ranks each entity in a dataset for each day in order of the field specified. Rows with equal values for the ranking criteria receive the same rank.	User-defined input field
max	Calculates the maximum value of a field over the specified timeframe.	Numerical fields
min	Calculates the minimum value of a field over the specified timeframe.	Numerical fields
ntile	Sorts the daily set of entities by the specified field and divides the list of entities equally into a specified number of buckets.	User-defined input field. Buckets: 1-100

rank	Ranks each entity in a dataset for each day in order of the field specified. Fields with equal values for the ranking criteria receive the same rank.	User-defined input field
row_number	Returns the sequential number of a row for each day in order of the field specified. Ties are broken by ENTITY NAME	User-defined input field
stddev	Calculates the standard deviation of the values over the day. Nulls are excluded.	Numerical fields
strength	Calculates an indicator that incorporates a decay function to give more weight to recent content. Applies only to numerical fields.	Numerical fields
sum	Calculates the sum of all the values for a field for each day.	Numerical fields

Codes:

Data Gaterhing .ipynb

Please double click below text box to see the all codes

class StockData

```
import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup
import yfinance as yf
import re
import io
import zipfile
import matplotlib.pyplot as plt
import statsmodels.api as sm
import random
from scipy import optimize as opt
from datetime import datetime


start_date = '2007-01-01'
end_date = '2022-12-31'



data = pd.read_csv("Ticker_US500.csv")
data = data['x'].tolist()
selected_sample = np.random.choice(data, 250, replace=False)
selected_sample = list(selected_sample)
tickers_to_remove = ['MS',
'DRE','CERN','BLL','ANTM','SIVB','NLOK','CTXS','PBCT','DISCK','DISCA','WTW','SNPS','DXC']
tickers = [t for t in selected_sample if t not in tickers_to_remove]
```

class PortfolioAnalysis

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from scipy import stats
import edhec_risk_kit_206 as erk
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm

liquidity_factors = pd.read_csv('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/liquidity_factors.csv',index_col=0)
book_to_market = pd.read_csv('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/book_to_market.csv',index_col=0)
market_cap_data = pd.read_csv('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/market_cap_data.csv',index_col=0)
stock_nominals = pd.read_csv('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/stock_nominals.csv',index_col=0)
factors_rf = pd.read_csv('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/factors_rf.csv',index_col=0)
stock_prices = pd.read_csv('C:/Users/toghruln/Desktop/FSV Thesis/Thesis_Data/stock_prices.csv',
index_col=0)
sentiment = pd.read_excel('C:/Users/toghruln/Desktop/FSV
Thesis/Thesis_Data/Sentiment_Sample_data.xlsx', index_col=0)


class PortfolioAnalysis:
    def __init__(self, stock_prices, liquidity_factors, book_to_market, market_cap_data, stock_nominals,
factors_rf):
        self.stock_prices = stock_prices
        self.liquidity_factors = liquidity_factors
        self.book_to_market = book_to_market
        self.market_cap_data = market_cap_data
        self.stock_nominals = stock_nominals
```

The list of selected tickers with the names of companies

No	Ticker	Name
1	PG	Procter & Gamble
2	EMN	Eastman Chemical
3	WLTW	Willis Towers Watson
4	HRL	Hormel Foods Corp.
5	ROP	Roper Technologies
6	BKNG	Booking Holdings Inc
7	MLM	Martin Marietta Materials
8	APH	Amphenol Corp
9	RTX	Raytheon Technologies
10	MO	Altria Group Inc
11	CMS	CMS Energy
12	KO	Coca-Cola Company
13	MMM	3M Company
14	ED	Consolidated Edison
15	DISH	Dish Network
16	PNR	Pentair plc
17	SBAC	SBA Communications
18	EOG	EOG Resources
19	EVRG	Evergy
20	DOV	Dover Corporation
21	HWM	Howmet Aerospace
22	A	Agilent Technologies
23	MPWR	Monolithic Power Systems
24	PLD	Prologis
25	WST	West Pharmaceutical Services
26	WMB	Williams Companies
27	ESS	Essex Property Trust, Inc.
28	LOW	Lowe's Cos.
29	LRCX	Lam Research
30	ETR	Entergy Corp.
31	LH	Laboratory Corp. of America Holding
32	LUV	Southwest Airlines
33	WDC	Western Digital
34	MSI	Motorola Solutions Inc.
35	TYL	Tyler Technologies
36	SPGI	S&P Global Inc.
37	WM	Waste Management Inc.
38	IEX	IDEX Corporation
39	XOM	Exxon Mobil Corp.
40	AZO	AutoZone Inc
41	ABT	Abbott Laboratories
42	WYNN	Wynn Resorts Ltd
43	ADBE	Adobe Inc.
44	BIIB	Biogen Inc.
45	WRB	W. R. Berkley Corporation
46	AFL	Aflac
47	NOV	NOV Inc.
48	AMZN	Amazon.com Inc.
49	RF	Regions Financial Corp.
50	HOLX	Hologic
51	O	Realty Income Corporation
52	WELL	Welltower Inc.
53	AKAM	Akamai Technologies
54	AWK	American Water Works
55	IVZ	Invesco Ltd.
56	ANTM	Anthem
57	FCX	Freeport-McMoRan Inc.
58	SHW	Sherwin-Williams
59	EW	Edwards Lifesciences
60	DIS	The Walt Disney Company
61	JPM	JPMorgan Chase & Co.
62	ADSK	Autodesk Inc.
63	CTSH	Cognizant Technology Solutions
64	DXC	DXC Technology
65	DTE	DTE Energy Co.
66	STE	Steris
67	OXY	Occidental Petroleum
68	USB	U.S. Bancorp
69	EIX	Edison Int'l
70	INCY	Incyte
71	CRM	Salesforce.com
72	BIO	Bio-Rad Laboratories
73	MCK	McKesson Corp.
74	J	Jacobs Engineering Group
75	WAB	Westinghouse Air Brake Technologies Corp
76	NTRS	Northern Trust Corp.
77	HBI	Hanesbrands Inc
78	EL	Estée Lauder Companies
79	MHK	Mohawk Industries
80	RHI	Robert Half International
81	BBY	Best Buy Co. Inc.
82	TRMB	Trimble Inc.
83	ADM	Archer-Daniels-Midland Co
84	MS	Morgan Stanley
85	VIAC	ViacomCBS
86	YUM	Yum! Brands Inc
87	ALB	Albemarle Corporation
88	DVA	DaVita Inc.
89	CNC	Centene Corporation
90	GD	General Dynamics
91	CAG	Conagra Brands
92	NUE	Nucor Corp.
93	EMR	Emerson Electric Company
94	CME	CME Group Inc.
95	ZION	Zions Bancorp
96	ARE	Alexandria Real Estate Equities
97	LB	L Brands Inc.
98	KLAC	KLA Corporation
99	DD	DuPont de Nemours Inc
100	PRU	Prudential Financial
101	BA	Boeing Company
102	TFX	Teleflex
103	MMC	Marsh & McLennan
104	MSCI	MSCI Inc
105	INTU	Intuit Inc.
106	ATVI	Activision Blizzard
107	ZBRA	Zebra Technologies
108	XLNX	Xilinx
109	FMC	FMC Corporation
110	SIVB	SVB Financial
111	GPN	Global Payments Inc.
112	AES	AES Corp
113	EFX	Equifax Inc.
114	ALL	Allstate Corp
115	PFE	Pfizer Inc.
116	NEE	NextEra Energy
117	PH	Parker-Hannifin
118	VRTX	Vertex Pharmaceuticals Inc
119	NWL	Newell Brands
120	LKQ	LKQ Corporation
121	RMD	ResMed
122	SLB	Schlumberger Ltd.
123	ANSS	ANSYS, Inc.
124	DISCK	Discovery, Inc. (Series C)
125	MGM	MGM Resorts International
126	BKR	Baker Hughes Co
127	XEL	Xcel Energy Inc
128	MKTX	MarketAxess
129	HAL	Halliburton Co.
130	DPZ	Domino's Pizza
131	HST	Host Hotels & Resorts
132	HAS	Hasbro Inc.
133	FDX	FedEx Corporation
134	TT	Trane Technologies plc
135	LIN	Linde plc
136	MNST	Monster Beverage
137	BMY	Bristol-Myers Squibb
138	DLTR	Dollar Tree
139	MXIM	Maxim Integrated Products
140	KIM	Kimco Realty
141	LHX	L3Harris Technologies
142	MOS	The Mosaic Company
143	TMUS	T-Mobile US
144	ABC	AmerisourceBergen
145	GE	General Electric
146	GPC	Genuine Parts
147	COST	Costco Wholesale Corp.
148	VFC	VF Corporation
149	MTB	M&T Bank
150	VLO	Valero Energy
151	GPS	Gap Inc.
152	PKI	PerkinElmer
153	EXPE	Expedia Group
154	FAST	Fastenal Co
155	MCD	McDonald's Corp.
156	TDY	Teledyne Technologies
157	AAPL	Apple Inc.
158	NI	NiSource Inc.
159	RE	Everest Re Group Ltd.
160	GILD	Gilead Sciences
161	PVH	PVH Corp.
162	TSCO	Tractor Supply Company
163	POOL	Pool Corporation
164	CSCO	Cisco Systems
165	GLW	Corning Inc.
166	AEP	American Electric Power
167	DUK	Duke Energy
168	BAC	Bank of America Corp
169	ZBH	Zimmer Biomet
170	VMC	Vulcan Materials
171	ISRG	Intuitive Surgical Inc.
172	FLIR	FLIR Systems
173	PNW	Pinnacle West Capital
174	ORCL	Oracle Corp.
175	T	AT&T Inc.
176	DRI	Darden Restaurants
177	NVR	NVR, Inc.
178	MRK	Merck & Co.
179	TSN	Tyson Foods
180	AEE	Ameren Corp
181	UDR	UDR, Inc.
182	SNA	Snap-on
183	PKG	Packaging Corporation of America
184	AJG	Arthur J. Gallagher & Co.
185	DRE	Duke Realty Corp
186	VRSN	Verisign Inc.
187	K	Kellogg Co.
188	TEL	TE Connectivity Ltd.
189	CDNS	Cadence Design Systems
190	STZ	Constellation Brands
191	EQR	Equity Residential
192	GIS	General Mills
193	IP	International Paper
194	GOOGL	Alphabet Inc. (Class A)
195	OMC	Omnicom Group
196	CINF	Cincinnati Financial
197	DHI	D. R. Horton
198	CMI	Cummins Inc.
199	AON	Aon plc
200	ABMD	Abiomed
201	MA	Mastercard Inc.
202	BEN	Franklin Resources
203	PEP	PepsiCo Inc.
204	HUM	Humana Inc.
205	WAT	Waters Corporation
206	PRGO	Perrigo
207	FISV	Fiserv Inc
208	FITB	Fifth Third Bancorp
209	AAL	American Airlines Group
210	PWR	Quanta Services Inc.
211	CBRE	CBRE Group
212	SCHW	Charles Schwab Corporation
213	GS	Goldman Sachs Group
214	HD	Home Depot
215	SRE	Sempra Energy
216	VTR	Ventas Inc
217	EXC	Exelon Corp.
218	MKC	McCormick & Co.
219	TXN	Texas Instruments
220	AOS	A.O. Smith Corp
221	VAR	Varian Inc.
222	HON	Honeywell Int'l Inc.
223	TPR	Tapestry, Inc.
224	ORLY	O'Reilly Automotive
225	ODFL	Old Dominion Freight Line
226	MAR	Marriott International
227	L	Loews Corp.
228	EBAY	eBay Inc.
229	RL	Ralph Lauren Corporation
230	GOOG	Alphabet Inc. (Class C)
231	UNH	UnitedHealth Group Inc.
232	PCAR	Paccar
233	KMB	Kimberly-Clark
234	LUMN	Lumen Technologies
235	CAH	Cardinal Health Inc.
236	LVS	Las Vegas Sands
237	ATO	Atmos Energy
238	PGR	Progressive Corp.
239	CF	CF Industries Holdings Inc
240	FRT	Federal Realty Investment Trust
241	EXPD	Expeditors
242	UNM	Unum Group
243	LNC	Lincoln National
244	CMA	Comerica Inc.
245	CTXS	Citrix Systems
246	TER	Teradyne
247	MDT	Medtronic plc
248	UAL	United Airlines Holdings
249	TTWO	Take-Two Interactive
250	NLOK	NortonLifeLock

Appendix II:

(Table: Authors)

Core Bibliography:

1. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263-291.
2. Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305-360.
3. Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2019). Cognitive neuroscience: The biology of the mind. W. W. Norton & Company.
4. By Richard L. Byyny, MD, FACP (https://www.med.upenn.edu/inclusion-and-diversity/assets/user-content/cognitive-bias.pdf)
5. Simon, H. A. (1991). Bounded rationality and organizational learning. Organization science, 2(1), 125-134.
6. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.
7. Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of Economic Perspectives, 17(1), 59-82.
8. DK. (2012). The Economics Book: Big Ideas Simply Explained. DK Publishing.
9. Seyhun, H. N. (1998). Investment Intelligence from Insider Trading. MIT Press Books
10. Jeng, L. A., Metrick, A., & Zeckhauser, R. J. (1999). The profits to insider trading: a performance-evaluation perspective. Harvard Institute of Economic Research Discussion Paper, (1880).
11. Fama, E. F. (1965). The behavior of stock-market prices. The Journal of Business, 38(1), 34-105.
12. Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. International Economic Review, 10(1), 1-21.
13. Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. Journal of Economic Perspectives, 17(1), 83-104.
14. Muhammad S. Tahir, Daniel W. Richards, Abdullahi D. Ahmed The role of financial risk-taking attitude in personal finances and consumer satisfaction: evidence from Australia, International Journal of Bank Marketing, 10.1108/IJBM-09-2022-0431.
15. Thaler, R. (1999). Mental accounting matters. Journal of Behavioral Decision Making, 12(3), 183-206.
16. Thaler, R. H. (1999). The end of behavioral finance. Financial Analysts Journal, 55(6), 12-17.
17. Elton, E. J., Gruber, M. J., & Padberg, M. W. (1978). Simple criteria for optimal portfolio selection. The Journal of Finance, 33(3), 583-603.
18. Capital Asset Pricing Model: Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3), 425-442.
19. Ross, Stephen A. (1976). "Arbitrage theory of capital asset pricing." Journal of Economic Theory 13 (3): 341–360.
20. Roll, Richard, and Stephen A. Ross. (1976). "The arbitrage pricing theory approach to strategic portfolio planning." Journal of Finance 31 (4): 1375–1387.

21. Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. The Journal of Finance, 47(2), 427-465.
22. Sharpe, W. F. (1991). The arithmetic of active management. Financial Analysts Journal, 47(1), 7-9.
23. Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. Review of Financial Studies, 29(1), 5-68.
24. Lo, A. W. (2017). Big data and new knowledge in finance. Journal of Economic Perspectives, 31(1), 3-28.
25. Li, X., Wang, Y., & Zhang, J. (2019). A news-based sentiment factor and its application to the stock market. Journal of Banking & Finance, 108, 105647.
26. Krishnamachari, R., & Tatonetti, N. P. (2019). A practical guide to building a financial market sentiment factor. Journal of Portfolio Management, 45(7), 105-117.
27. Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. Handbook of the Economics of Finance, 1, 1053-1128.
28. Odean, T. (1998). Are investors reluctant to realize their losses?. The Journal of Finance, 53(5), 1775-1798
29. Pompian, M. M. (2006). Behavioral finance and wealth management: How to build optimal portfolios that account for investor biases. John Wiley & Sons.
30. Fagerström, A. (2008). Overconfidence and over positivity in the stock market. Journal of Behavioral Finance, 9(1), 1-12.
31. Balcilar, M., Gupta, R., Wohar, M.E. (2021). "The effects of the COVID-19 pandemic on the U.S. economy: Evidence from monthly macroeconomic indicators." Journal of Macroeconomics, 67, 103312.
32. Bae, K., Lee, J., & Park, K. (2020). How Does COVID-19 Affect Corporate Financial Performance?. Available at SSRN 3620093.
33. Chen, Z., & Zheng, S. (2020). Artificial intelligence in finance: A review. Journal of Finance and Data Science, 6(3), 239-254.
34. Gomber, P., Koch, J., & Siering, M. (2018). High-frequency trading and its role in fragmented markets. Journal of Business Research, 88, 204-218.
35. Hu, Y., & Choudhary, V. (2019). Artificial intelligence in financial services: Emerging applications and regulatory concerns. Journal of Financial Regulation and Compliance, 27(3), 365-379.
36. Makridakis, S., & Taleb, N. (2019). Decision making and artificial intelligence in the financial industry. International Journal of Forecasting, 35(3), 1025-1039.
37. Beck, R., & Storkenmaier, A. (2018). Machine learning in finance: A state-of-the-art survey. Applied Economics, 50(3), 266-277.
38. Nacher, G., Kornprobst, P., & Alet, P. J. (2020). Deep learning and finance: A review. Journal of Risk and Financial Management, 13(8), 172.
39. Zhang, W., Wang, Y., & Luo, X. (2019). A deep learning framework for stock sentiment analysis based on financial news. Future Generation Computer Systems, 94, 242-249.
40. Tetlock, P. (2007). “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” Journal of Finance, Vol. 62, No. 3, pp.1139–1168. Available at:
41. Baker, M., Ruback, R. S., & Wurgler, J. (2007). Behavioral corporate finance: A survey. Foundations and Trends in Finance, 1(1), 1-101

42. Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129-151.
43. Baker, M., & Wurgler, J. (2013). Do investors vote with their feet? Journal of Financial Economics, 107(2), 478-498.
44. 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
45. Zaremba, A., & Bystrzycka, E. (2018). The performance of MSCI Barra Global Equity Model 3 in emerging markets. Contemporary Economics, 12(4), 395-412. doi:10.5709/ce.1897-9254.267
46. "Natural Language Processing with Python" by Steven Bird, Ewan Klein, and Edward Loper.
47. Islam, M. A., Ashraf, M. N., Abir, A. H., & Mottalib, M. A. (2018). An approach to classify online news based on sentiment analysis performed at the sentence level. International Journal of Electrical and Computer Engineering (IJECE), 8(1), 307-315
48. Marneffe, M.C. and Potts, C. (2005). "Part-of-Speech Tagging". In: Handbook of Natural Language Processing. Edited by N. Indurkhya and F. J. Damerau. CRC Press. pp. 141-160.
49.
50. Ben-David Et Al -(Unsupervised Domain Adaptation Based on Source-guided Discrepancy).
51. Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. Journal of the American Statistical Association 101(473):138–156.
52. Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. In S.Makridakis, The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms (pp. 46-60). Elsevier
53. Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. Journal of Political Economy, 100(5), 992-1026.
54. Banerjee, A. V. (1992). A Simple Model of Herd Behavior. The Quarterly Journal of Economics, 107(3), 797-817.
55. Angela Maria Filip, Maria Miruna Pochea(2023) “Intentional and spurious herding behavior: A sentiment driven analysis”