Early-Stage Venture Financing: A Data-Driven Approach with Machine Learning Application

Pornpanit Rasivisuth

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Institute of Finance and Technology
Department of Civil and Environmental Engineering
University College London

July 3, 2025

I, Pornpanit Rasivisuth, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Venture capital (VC) and private equity (PE) become indispensable financial assets, globally driving economic and societal growth. This project primarily aims to investigate the utility of alternative datasets and machine learning models in addressing various challenges prevalent within private markets. These challenges, often exacerbated by the illiquid nature of private investments, information asymmetry, and potential moral hazard issues, include the valuation of Initial Coin Offering (ICO) tokens, the assessment of early-stage company valuations, and the selection of startups for venture capital funds. A Natural Language Processing (NLP) model, capable of analysing unstructured text data, is employed alongside additional signals derived from alternative data sources such as social media and financial news. Overall, the project is anticipated to benefit academic researchers and practitioners within the private capital sectors, contributing to recent advancements in both the technology and sustainable finance domains.

Impact Statement

Venture capital, a cornerstone of the private capital market, focuses on investing in and supporting the growth of early-stage companies characterised by higher risk profiles, illiquidity, and long-term investment horizons. However, the process of assessing these entrepreneurial ventures is hindered by information asymmetries, moral hazard, and the limitations of traditional datasets. Moreover, the evolving landscape of technology, such as blockchain, and the growing emphasis on sustainable finance necessitate a re-evaluation of investment strategies. To address these challenges, this research introduces a novel approach that leverages alternative datasets and machine learning to enhance the screening and due diligence processes during the initial stages of private capital investment.

By addressing the shortcomings of existing token rating methodologies and whitepaper analysis used in token financing for initial coin offerings (ICOs), the first paper introduces an ML-based framework that leverages social media sentiment to predict token returns. The findings underscore the critical role of understanding market sentiment and investor behaviour through social media in shaping token return predictions and highlight the systemic risks inherent in the lightly regulated ICO market. These insights not only advance academic knowledge but also provide valuable guidance for practitioners and policymakers seeking to foster a more transparent and investor-centric token economy.

The second study introduces a pioneering valuation framework for early-stage companies that surpasses traditional financial analysis by seamlessly integrating machine learning and sustainability data. Incorporating ESG indicators, the model provides a more comprehensive and forward-looking assessment of early-stage companies, demonstrating the value of alternative, unstructured datasets. This approach not only enhances investment decision-making but also aligns with growing investor demand for sustainable investments. Addressing the limitations of traditional valuation methods like discounted cash flow, which often rely on inaccessible financial statements, the study unveils the predictive power of a novel model to capture the complexities of early-stage companies. This significantly advances valuation theory and practice, fostering a new generation of valuation models that are robust, transparent, and aligned with long-term value creation and sustainability regulations.

Lastly, the final study utilises reinforcement learning (RL), an underexplored area in private capital investment, to revolutionise venture capital investment decision-making. By developing a novel RL-based recommendation system, the study introduces a paradigm shift in portfolio management, enabling investors to identify high-potential startups and optimise investment returns. The research explores the intricate design choices inherent in RL, including state representations, reward functions, and exploration strategies, tailoring these elements to the unique challenges of the venture capital landscape, especially the illiquidity nature. This research profoundly impacts both academics and practitioners, encouraging future AI-driven investment research while offering a practical tool for venture capitalists seeking a competitive advantage.

Therefore, overall this thesis suggests that the alternative dataset can be used alongside commercial databases available in private capital markets to support the investment lifecycle in addition to the machine learning models. The same application can be further applied in private equity, where the focus is on mature companies.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Fiaschetti, and Professor Medda for their invaluable guidance and unwavering support throughout my PhD studies. I am also deeply thankful to the Institute of Finance and Technology for providing the funding that enabled me to pursue my research at the Department of Civil and Environmental Engineering, University College London. I am also grateful to Dr. Indjic for his guidance and the valuable perspectives he offered throughout my research.

This thesis could not have been completed without the unwavering support I received from my family and friends throughout this challenging academic journey. I owe an immense debt of gratitude to my parents, whose encouragement and support enabled me to pursue my academic journey in the United Kingdom. I must also thank my partner for his patience and for keeping me well-nourished. His cooking skills were critical to my survival—both mentally and physically—throughout this thesis journey.

A heartfelt thanks to my friends for the much-needed distractions, late-night conversations, and for reminding me that there's more to life than just research (even though I may not have believed you at the time). Special thanks also go to my trainers, who helped me discover a new sport hobby.

UCL Research Paper Declaration

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** An investigation of sentiment analysis of information disclosure during Initial Coin Offering (ICO) on the token return

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1016/j.irfa.2024.103437

   (c) **Where was the work published?** International Review of Financial Analysis

   (d) **Who published the work?** Elsevier

   (e) **When was the work published?** 10 July 2024

   (f) **List the manuscript's authors in the order they appear on the publication:** Pornpanit Rasivisuth, Maurizio Fiaschetti, Francesca Medda

   (g) **Was the work peer reviewd?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** No

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**
   If 'Yes', please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order: Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** The author P.R. confirmed responsibility for the following: study conception and design, data collection, model, implementation, analysis and interpretation of results, and manuscript preparation. All authors contributed to reviewing and editing the final version of the manuscript. M.F. and F.M. supervised the project.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date: 02/05/2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 02/05/2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Startups, characterised by their innovation-driven nature, are entrepreneurial ventures that leverage financial, human, and other resources to introduce novel products or processes into the marketplace. These ventures often involve technological advancements, intensive research and development, and innovative business models (El Hanchi and Kerzazi, 2020). For startups to truly drive significant economic development, innovation, and social progress, they critically depend on financial resources and strategic guidance, particularly from private capital in exchange for equity.

Private capital, including venture capital and private equity, serves as a critical source of funding and support for early-stage companies. Beyond financial resources, VC and PE firms often provide strategic guidance, industry expertise, and access to valuable networks, fostering value creation and maximising returns upon exit. While both VC and PE invest in early-stage companies, they differ primarily in their investment focus. VCs typically invest in early-stage startups, which may not yet generate revenue or enter the market. This type of investment carries higher risk due to the significant uncertainty surrounding potential returns and the possibility of negative cash flows during the valley of death phase. In contrast, stable firms are often financed through growth equity or leveraged buyouts, which can involve both equity and debt, commonly known as private debt.

Despite the inherent risks and illiquidity associated with private market investments, the potential for extreme returns makes them an attractive asset class for both individual and institutional investors seeking to diversify their portfolios. Given the higher risk associated with early-stage investing, there is a pressing need for further research to understand the unique characteristics of startup investments and the underlying challenges confronting both investors and entrepreneurs in this space.

The challenges found in entrepreneurial finance are characterised by two additional issues: asymmetric information and moral hazard (Denis, 2004) which affect two key stakeholders in the ecosystem - investors and entrepreneurs. Information asymmetry describes the phenomenon when the party with an information advantage make a better decision than another without enough information. Investors encounter difficulties in valuing the company given the limited information, especially for startups with a lack of cash flows, collateral, and financial statements (Damodaran, 2009; Sander and Kõomägi, 2007); as a result, entrepreneurs themselves suffer these constraints to raise external funding (Block et al., 2018). Another problem is the moral hazard in which entrepreneurs may misallocate the funding; the term sheet can resolve this issue to have agreement on economics and governance interest (Denis, 2004) or gather information to monitor the progress of the development (Wang and Zhou, 2004). These two fundamental issues lead to the emergence of entrepreneurial finance in various directions, including the entry of new players leveraging Blockchain technology, the emergence of the sustainability investment trend, and the shift toward digitalisation involving the integration of big data and artificial intelligence (AI).

The landscape of private capital investment is significantly reshaped by the emergence of dedicated databases such as Pitchbook, Preqin, and Crunchbase. These platforms empower both academics and practitioners to conduct data-driven analyses of target portfolio companies and their respective industries. However, a critical challenge persists: the absence of specific regulations governing private capital data often impedes robust validation processes, which traditionally rely on investor due diligence. This inherent confidentiality consequently limits public access to verified startup information. To bridge this data gap, alternative data sources become increasingly important. These sources, often unstructured and distinct from conventional financial statements, encompass diverse formats, including text, images, and videos

(Cong et al., 2021). Their capacity to enhance financial decision-making and reduce data acquisition costs garners considerable attention from both academic researchers and industry practitioners.

Concurrently, machine learning (ML), a subset of AI capable of learning from data without explicit programming, emerges as an exceptionally valuable tool within the financial sector. ML applications span diverse financial domains, from retail and investment banking to payments, insurance, and critically, private capital investment, where they can significantly assist in screening and valuation processes (Miloud et al., 2012; Ang et al., 2022; Zhang et al., 2023; Garkavenko et al., 2021). This thesis systematically assesses the extent to which advanced data analytics and machine learning can be integrated to support private capital investments, particularly at the early stages of venture capital financing. This study delves into three compelling research frontiers, which form the core of its empirical chapters: (1) token financing through Initial Coin Offerings (ICOs), (2) the sustainability transition within private capital investment, and (3) leveraging advanced machine learning models for screening investment opportunities.

The first study (Chapter 3) is titled "An investigation of sentiment analysis of information disclosure during Initial Coin Offering (ICO) on the token return". The emergence of blockchain technology offers an alternative to traditional entrepreneurial financing, enabling capital raising through token finance. However, this new paradigm presents significant valuation challenges due to information asymmetry and a lack of regulatory oversight, despite the availability of the token rating platforms (Ofir and Sadeh, 2020; Florysiak and Schandlbauer, 2019). This research's objective is to determine if token ratings guarantee positive ICO returns and to analyse what alternative data (whitepapers and social media) can impact long-term token returns.

To achieve this, the research constructs an ICO index coupled with sentiment analysis from whitepapers and Twitter (currently known as X) and implements machine learning to predict token returns utilising available features and alternative data. The results confirm a discrepancy between ICO ratings and actual returns, highlighting the unreliability of current rating systems and underscoring the crucial role of social media and machine learning in overcoming information asymmetries. This chapter contributes significantly by empirically evidencing the limitations of existing ICO rating systems, developing a methodological approach using integrated alternative data and sentiment analysis, and demonstrating the potential of machine learning to enhance transparency and predictive accuracy in the valuation of token financing.

The next chapter (Chapter 4) titled “Startup Valuation with Sustainability: A Novel Approach with Machine Learning and Natural Language Processing” emphasises the impact investment as a prevalent trend driven by the public interest and the governance policies that encourage the companies to disclose their environmental, social, and governance (ESG) performance. Reflecting this trend, many investors in publicly traded equities integrate ESG-related information into their investment strategies, leading to extensive research on the relationship between ESG data and asset pricing in public markets (Serafeim and Yoon, 2023, 2022; Shanaev and Ghimire, 2022; Gibson Brandon et al., 2021). Responding to calls from researchers like Cumming et al. (2022), this study’s research objective is to investigate the unexplored impact of sustainability investment on early-stage startup valuation. This is particularly critical given the limited availability of traditional sustainability data and the consistency issues in existing ESG ratings (Boffo and Patalano, 2020).

To address these challenges, the study proposes a novel approach: integrating natural language processing (NLP) and machine learning (ML) to extract sustainability indicators from startup news, which serve as an alternative data source to enhance the pre-money valuation of early-stage companies. The study introduces a novel ML-based valuation model that conceptually advances traditional methodologies by investigating the impact of sustainability-related textual data on company valuation. The results indicate a substantial 16.45% improvement in prediction accuracy over conventional approaches, demonstrating the significant potential of sustainability context to enhance startup valuation alongside traditional characteristics and funding round data. This contribution uniquely bridges the domains of sustainable finance and quantitative valuation, offering practitioners in the private capital market an innovative framework that considers sustainability’s crucial long-term impact alongside short-term financial gains.

Chapter 5, “Identifying High-Growth Startups: A Reinforcement Learning Approach for Venture Capital," introduces a novel application of reinforcement learning (RL) in private capital. While supervised models are prevalent in finance, RL remains underexposed in this domain, despite its proven ability in recommender systems (RLRS) to surpass traditional methods by enabling agents to interact with their environment (Chen et al., 2019; Taghipour et al., 2007; Liebman and Stone, 2014; Lei and Li, 2019). This chapter is motivated by the potential of RL to address information asymmetry in private capital by recommending early-stage startups for portfolio construction, particularly given the limited exposure of RL in this sector compared to supervised learning.

This study aims to propose and evaluate a new RLRS model, called VC-RLRS, specifically designed for venture capital investment. This involves recommending top-ten portfolio companies and outlining how different configurations impact recommendation performance. The model is built upon Q-learning, integrating the unique limitations of VC investment directly into the design and choices of its state representations and reward functions. The VC-RLRS demonstrates a strong capability to recommend high-growth startups, explicitly accounting for crucial factors like exit opportunities and portfolio diversification. This not only conceptually extends recommender systems to complex financial ecosystems but also showcases their potential to enhance investment decision-making for both generalist and specialist strategies, with successful applications across FinTech, Healthcare, and Information Technology. The study further contributes significantly by integrating deep learning into a hybrid model to assess performance against a Q-learning baseline, identifying areas for future scalability. Overall, this chapter offers an original design and evaluation of RL components tailored for the VC context, substantially advancing the field and outlining promising avenues for practical VC application.

My thesis is composed of a literature review outlining the overview of venture capital and private equity investment, and the application of alternative datasets and machine learning models in private capital investment. This is followed by the three themed chapters that are previously described. The final chapter (Chapter 6) presents the conclusion of this thesis, alongside an overview of contributions, a discussion of limitations and future research directions.

Chapter 2

Literature review

Venture capital (VC) and private equity (PE) become indispensable financial asset classes that contribute to and drive economic growth and society globally. The assets under management (AUM) within the global private capital market exhibit sustained growth, reaching \$9.8 trillion in 2021 (McKinsey, 2022). Numerous successful public companies and unicorns (those valued at over \$1 billion) benefit from the strategic investments of VC and PE firms. A prime example is Meta (formerly Facebook), a social media platform, which secured Series A funding from Accel Partners in 2005, valued at \$98 million (Pitchbook, 2022). Through subsequent growth and a successful IPO on the Nasdaq stock exchange in 2012, Meta's valuation surged to \$81.25 billion, yielding Accel Partners a substantial return on their initial investment from its ownership of \$6.3 billion (Pitchbook, 2022; Tam and Raice, 2012; McBride, 2012). Accel Partners' strategic guidance and support play a pivotal role in Meta's growth, demonstrating the value of experienced VC firms in accelerating high-potential startups. This case study underscores the pivotal role of private capital in fostering the growth of high-potential startups.

Beyond the well-known unicorn startups that go public through initial public offerings (IPOs), numerous other companies previously backed by VC and PE firms achieve significant milestones. For instance, Slack, a business communication platform, was acquired by Salesforce in 2021 after receiving funding from Accel Partners (Salesforce, 2021). Additionally, Gocardless, a recurring payment collection platform founded in 2011, continues to operate privately with ongoing support and follow-on investment from Accel (Pitchbook, 2024). Collectively, VC-backed companies raised a substantial $144.8 billion globally in Q1 (KPMG, 2022), underscoring their significant impact on the startup ecosystem in terms of financial funding.

Venture capital and private equity investments play a pivotal role in nurturing the growth of early-stage companies, which in turn contribute significantly to economic development, innovation, and social progress. In the United States alone, startups and young firms generated over 3.7 million new jobs in 2023 (Statista Research Department, 2024). Moreover, startups comprise approximately 20% of the workforce across OECD countries (OECD, 2024). Government policies also play a crucial role in shaping the dynamics of startups and job creation, influencing the overall economic landscape (Calvino et al., 2016; Kane, 2010). Beyond their direct economic contributions, startups supported by private capital often contribute to social and environmental progress by addressing pressing challenges and developing sustainable solutions.

As evidenced by the number of patents obtained, VC-backed companies often demonstrate a higher propensity for innovation (Nanda and Rhodes-Kropf, 2013). In Addition, startups focus increasingly on addressing pressing global challenges. Climatetech startups attract significant investment, reaching a valuation of $2.5 trillion (Dealroom, 2024a). Meanwhile, the Healthtech startups demonstrate a strong commitment to improving quality of life, with over 120,000 patent activities recorded between 2010 and 2021 and $25 billion invested globally in 2023 (Dealroom, 2023). The proliferation of startups across various sectors positively impacts sustainable economic development and social progress, as evidenced by their strong correlation with the United Nations Sustainable Development Goals (UN SDGs) (Ressin, 2022).

Building on the discussion of private capital's impact on startups, the economy, and society, this chapter provides a comparative overview of early-stage investment, specifically focusing on venture capital and private equity. It then explores the transformative potential of alternative datasets and machine learning within the financial market, critically highlighting their currently limited application to private capital investment.

2.1 The background of venture capital and private equity

These firms invest in private companies through equity and debt financing, aiming to generate returns from valuation growth. While both seek to capitalise on private company potential, a key distinction lies in their investment stage. VC firms focus on early-stage, high-risk startups often lacking revenue, whereas PE firms target more mature companies seeking private growth capital (i.e., prior to initial public offering or IPO). According to a KPMG report, the average deal size for early-stage ventures is smaller at $7.9 million compared to $13.5 million for later-stage investments; however, the number of early-stage deals is substantially higher (KPMG, 2022). The investment process for both typically involves formulating an investment strategy, fundraising from limited partners (LPs), and selecting portfolio companies aligned with the investment objectives. Comprehensive due diligence is conducted on potential investments, considering factors such as the founding team, product, market, and legal aspects. This process is crucial for identifying promising opportunities and mitigating investment risks. The subsequent phase involves deploying the capital committed by LPs.

Beyond monetary investment, VC and PE firms might offer invaluable strategic counsel and access to their extensive networks, significantly contributing to the growth and success of portfolio companies and facilitating exits like M&A or IPOs. This support encompasses talent recruitment, customer and supplier introductions, and other value-added services. The funding landscape evolves significantly with the emergence of new players like angel investors, crowdfunding, and corporate venture capital (CVC), intensifying competition for investment opportunities among startups (Block et al., 2018). Angel investors are high-net-worth individuals who provide early-stage funding to startups in exchange for equity ownership or convertible debt. Crowdfunding, in contrast, allows a large number of individuals to collectively finance ventures through digital platforms. The CVC refers to investments made by established corporations in startups to achieve strategic advantages alongside financial returns. Additionally, technological advancements such as decentralised blockchain networks pave the way for innovative financing approaches, giving rise to new categories of startups and investment strategies that are explored in greater detail in Chapter 3.

Private capital can be defined as an alternative asset class, that offers investors the opportunity to diversify their portfolios and mitigate risk. Investors can choose to invest in these private capital funds managed by general partners (GPS), who are entities responsible for overseeing venture capital funds, or they can invest directly in portfolio companies, avoiding management fees and agency problems. Markowitz (1952)'s Modern Portfolio Theory (MPT) introduces a framework for constructing optimal portfolios by balancing expected return and risk. However, the theory's underlying assumption of asset liquidity and market efficiency renders it inapplicable to private equity (Thomas and Pierre-Yves, 2005). Unlike traditional asset classes, private equity investments are characterised by their illiquid nature, requiring investors to commit capital for extended holding periods. LPs provide capital to the fund through capital calls and receive returns upon the exit of portfolio companies, which can take up to a decade. The return distribution of private equity investments deviates significantly from the normal distribution observed in traditional asset classes, exhibiting a heavily skewed distribution with a concentration of returns in a few highly successful investments (Cochrane, 2005). The distinctive return patterns of private equity investments necessitate a specialised investment strategy that prioritises identifying and nurturing high-potential opportunities, conducting due diligence, and actively monitoring portfolio companies to achieve exceptional returns that compensate for the inherent risks.

As previously discussed, information asymmetries and moral hazard issues remain prevalent challenges in entrepreneurial financing, hindering accurate valuation and capital allocation (Denis, 2004). The landscape of VC investing evolves significantly, characterised by a shifting focus on technology-driven startups, diminishing governance standards, and evolving decision-making criteria among investors (Lerner and Nanda, 2020). To navigate these challenges and enhance investment outcomes, investors are increasingly turning to data-driven approaches and advanced analytical tools to resolve issues in entrepreneurial finance. This leads to the development of key areas explored in this thesis: token financing rating for ICO investments (Chapter 3), the application of machine learning to perform valuation (Chapter 4), and its use in portfolio selection for early-stage startups (Chapter 5). The following sections explore the availability of relevant data sources for both investors and academic researchers, followed by an in-depth examination of machine learning applications as a promising tool for identifying target portfolio companies and optimising investment returns.

2.2 Private capital database and alternative data

2.2.1 Private capital database

Understanding the determinants of startup funding, valuation, and growth necessitates robust data. While public companies are obligated to disclose financial and non-financial information subject to regulatory scrutiny, private companies, particularly early-stage startups, have limited obligations and incentives to share such information. Consequently, data availability for these firms is limited and often proprietary posing significant challenges for researchers and investors. Additionally, unlike public securities with frequent market data, the illiquid nature of private equity investments hinders the analysis of investment trends, valuation patterns, and performance metrics. The validation of startup information primarily relies on investor due diligence, with results often remaining confidential. However, commercial databases like PitchBook, Preqin, and Crunchbase offer valuable resources by aggregating data on companies, founders, deals, and investment funds, providing essential information for both academics and practitioners.

Several studies compare different private capital investment databases. For instance, Maats and Bedrijfskunde (2008) contrast the firm-level and funding round-level coverage of VentureXpert[1] and Venture Source[2]. While VentureXpert offers more comprehensive data, it also exhibits higher error rates. Kaplan and Lerner (2016)

[1] VentureXpert is part of Thomson Reuters

[2] Venture Source was formerly part of Dow Jones, acquired by CB Insights in 2020 (CB Insights updates, 2020)

emphasise the importance of accurate pre- and post-valuation of startups, highlighting the need for databases to reflect the control rights of stakeholders, including cash flow, voting, and liquidation rights, as outlined by Kaplan and Strömberg (2003), rather than relying solely on the simple multiplication of share price and quantity. However, without publicly accessible deal terms, accurately calculating valuations remains challenging. A broader range of databases, including Cambridge Associates, AngelList, Burgiss[3], CB Insights, Crunchbase, Dealroom, Pitchbook, Preqin, and Tracxn, are assessed by Kaplan and Lerner (2016) and Retterath and Braun (2020). Each database possesses unique characteristics, and researchers should be mindful of potential biases when utilising these data sources.

For example, Retterath and Braun (2020) find that startups in sectors such as IT, software, biotech, and healthcare are more likely to be included in databases like Crunchbase. PitchBook and Crunchbase tend to include younger companies, reflecting the time trends. Additionally, founders with prior funding successes or involvement in M&A deals are more prominently featured in PitchBook datasets. Retterath and Braun (2020) further confirm that later-stage funding rounds exhibit more complete data on round sizes and post-money valuations compared to early-stage investments. Their analysis also reveals that VentureSource, PitchBook, and Crunchbase offer the most comprehensive and accurate data across key dimensions of company information, founder data, and funding details. These findings underscore the persistent challenge of obtaining reliable early-stage funding data and addressing inherent biases within these datasets.

Kaplan and Lerner (2016) highlights that biases significantly distort data regarding the performance of VC funds. Several factors contribute to biases in VC fund performance data, including the underrepresentation of first-time funds and the limited transparency regarding underperforming funds. These factors can lead to an upward bias in reported performance due to the selective disclosure of information, as GPs may avoid reporting underperforming funds to protect their reputation. Additionally, the complex and non-transparent valuation methodologies, as detailed in Chapter

[3] Burgiss is now part of MSCI

4, employed by VCs can hinder benchmarking and comparative analysis. To address these limitations and enhance the analysis of private companies, practitioners and academic researchers can explore alternative datasets that can complement traditional commercial and structured data sources.

2.2.2 Alternative data in finance

Alternative data, encompassing information beyond traditional financial statements and company filings, becomes increasingly accessible due to technological advancements. This diverse dataset, including text, images, and audio, garners significant attention for its potential to enhance financial decision-making and reduce data acquisition costs from both academics and practitioners (Cong et al., 2021). Textual data, such as financial news, reports, and social media content, is a commonly used type of alternative data that can be collected through web scraping[4] and APIs [5] (Śpiewanowski et al., 2022; Cong et al., 2021). Other forms of alternative data used in financial applications include satellite imagery (Yang and Broby, 2020) and audio (Mayew and Venkatachalam, 2012). Although alternative data presents promising opportunities, it also poses significant technical challenges, particularly in processing, analysing, and storing unstructured information. Furthermore, data privacy concerns require careful attention from regulators to ensure its responsible and ethical use (Cong et al., 2021).

While numerous studies explore the application of alternative data in publicly traded equities, textual data remains the most widely used source. For example, sentiment analysis of Twitter data is used to predict stock price movements (Pagolu et al., 2016) and returns (Ranco et al., 2015). Moreover, web traffic metrics, including website interactions and user behaviour, are associated with the performance of internet stocks (Hand, 2001). Additionally, search engine queries related to portfolio companies are correlated with trading volume (Bordino et al., 2012). Financial news feeds constitute another popular source of alternative data, widely employed by investors and academics. Researchers explore various textual representation

[4] The process of extracting data from HTML pages and storing it in a structured format

[5] Application Programming Interface or API is a set of protocols that allows different software applications to communicate, interact, and exchange data with each other.

techniques to extract valuable insights from news articles (Schumaker and Chen, 2009; Mittermayer, 2004). By combining feature selection methods with market feedback, studies demonstrate the potential of these techniques to improve the accuracy of price movement prediction and trading strategy performance (Hagenau et al., 2013). Chapters 3 and 4 also highlight the importance of these alternative data sources, demonstrating the extensive use of social media and financial news in predicting token returns and startup valuations, respectively.

In the context of private capital, social media platforms like X (formerly known as Twitter) emerge as valuable predictors of venture success. Antretter et al. (2019) demonstrate the efficacy of online legitimacy, a measure of social appreciation on Twitter, in forecasting 5-year startup survivability with 76% accuracy. Bayar and Kesici (2024) further highlight the correlation between higher Twitter engagement levels and fewer VC financing rounds, smaller VC syndicates, and a greater likelihood of successful exits with larger funding amounts. Garkavenko et al. (2022) leverage social media and web-based metrics to predict startup funding success, emphasising the potential of publicly available sources to provide labelled target variables that may be absent in traditional databases.

Beyond social media, text-based analysis is applied to various types of data. Trevor Rogers (2020) find a positive correlation between patent similarity among VC-backed portfolio companies and their subsequent patent output and quality. While venture capitalists are prominent players in early-stage funding, other sources such as angel investors, crowdfunding, and initial coin offerings (ICOs) also gain traction. These alternative funding sources often leverage textual data from crowdsourcing platforms to assess investment opportunities. Studies show that direct mentions of startups in business and employment-related news articles can enhance the predictive power of models for angel and seed funding success (Sharchilev et al., 2018). Kaiser and Kuhn (2020) further explore the potential of publicly available data, including textual information from the Danish Business Authority, to predict various dimensions of Danish startup performance, such as survivability, employment growth, and return on assets. Additionally, Lee et al. (2021) investigate the impact of ICO aggregated ratings, a collective opinion from communities and experts, on fundraising success and long-run token returns. These ICO ratings are discussed further in Chapter 3.

The increasing prominence of sustainable finance leads to a growing interest in news articles and sustainability reports as valuable sources of alternative data. Studies demonstrate the impact of these text data on future asset returns and company risk (Guo et al., 2020; Schmidt, 2019). Satellite technology, which enables the monitoring of environmental indicators such as air and water pollution, waste management, and natural resource management, offers another valuable source of information aligned with EU legislation. This data can contribute significantly to ESG performance measurement for financial assets (Yang and Broby, 2020). Chapter 4 delves deeper into the application of alternative data in sustainable finance assessment.

Another application of alternative data involves the analysis of audio recordings, such as earnings conference calls. Mayew and Venkatachalam (2012) employ vocal emotion analysis software to extract emotional cues from these calls, demonstrating their potential to provide insights into a company's financial future. These emotional cues can complement traditional financial data and textual analysis. Collectively, these findings highlight the potential of alternative data, a relatively untapped potential resource in private capital investment. Such data can offer insights into the investment process, particularly during the screening and due diligence phases. The subsequent section delves deeper into the application of machine learning models to harness the capabilities of these alternative data sources

2.3 Application of machine learning in private capital market

Machine learning (ML), a subset of artificial intelligence that is capable of learning from various types of data without explicit programming, emerges as a valuable tool in financial markets, including asset management, insurance, and risk management. Its application in private capital investment is gaining traction, particularly in the areas of early-stage screening and due diligence. By judiciously selecting investment opportunities through machine learning, investors can significantly enhance the value creation of target startups (Gompers et al., 2020).

Identifying exceptional investment opportunities capable of generating superior returns remains a significant challenge for venture capital and private equity firms. Predictive models emerge as a potential solution, with research focusing on forecasting startup outcomes such as subsequent funding rounds (Sharchilev et al., 2018; Garkavenko et al., 2022), survivability (Krishna et al., 2016), and exit pathways (Bhat and Zaelit, 2011; Arroyo et al., 2019; Ross et al., 2021). Additionally, studies explore the application of machine learning to predict startup valuations (Miloud et al., 2012; Ang et al., 2022; Zhang et al., 2023; Garkavenko et al., 2021). The majority of these studies employ supervised learning, where algorithms are trained on labelled data with known target variables. Moreover, the application of supervised learning shows success in public equity investment, especially for stock prediction (Lawal et al., 2020; Kumar et al., 2018; Powell et al., 2008).

Unsupervised learning, another ML technique, trains algorithms on unlabelled data to identify underlying patterns and structures. Despite the absence of labelled training data, unsupervised learning models demonstrate comparable performance to supervised learning models in market forecasting tasks (Powell et al., 2008; Corchado et al., 1998). While it shows comparable performance in public equities research, its application in private capital research remains limited. Unsupervised learning is employed to understand relationships and classify startups into industry domains based on textual descriptions (Kharchenko et al., 2023). Additionally, semi-supervised learning, a hybrid approach combining supervised and unsupervised techniques, shows promise in both public and private equities research. In the context of public equities, this approach is applied to improve time series forecasting for Nasdaq markets (Palma et al., 2024) and spot foreign exchange rates (Pavlidis et al., 2006). In the realm of private capital, Xiong and Fan (2021) implement a semi-supervised approach to analyse VC network structures and identify industry leaders.

Reinforcement learning is an ML method where an agent learns to make decisions by interacting with an environment. It is successfully applied in liquid financial markets to optimise trading strategies, such as stock trading (Azhikodan et al., 2019) and asset allocation (Moody and Saffell, 2001). However, its application in the illiquid private capital market is limited due to extended investment horizons. Designing reinforcement learning models for this context is challenging due to delayed feedback mechanisms inherent to venture capital and private equity investments. Chapter 5 examines the potential applications of reinforcement learning within the private capital domain in more detail.

Finally, a neural network (NN), a computational model inspired by the human brain, learns complex patterns from data through iterative adjustments to minimise error. Information flows through the network, from input to output layers, undergoing transformations at intermediate hidden layers. Monika Dhochak and Doliya (2024) apply neural networks to predict pre-money valuations of Indian startups. Li et al. (2022) extend the NN model by utilising graph neural networks (GNNs) to predict M&A probabilities, leveraging the network relationships among companies, founders, and investors. Deep learning, on the other hand, is a specialised neural network architecture with multiple layers, that excels at extracting intricate features from complex data. Ross et al. (2021) employ deep learning to develop an ensemble model, CapitalVX, which outperforms human venture capitalists in predicting startup exit scenarios. Natural Language Processing (NLP) is another domain where deep learning shows promise in processing and analysing the unstructured text data format. For instance, Caragea et al. (2020) utilise deep learning and BERT transformers to categorise fintech innovations based on patent analysis. Chan et al. (2021) apply BERT to crowdfunding project descriptions, generating scores indicative of writing quality and predicting fundraising success. Interestingly, lower-quality writing given by a higher average BERT score correlated with increased funding.

As previously discussed, the application of alternative data and machine learning in private capital investment remains relatively limited compared to its use in public financial markets. The availability of robust datasets is essential for training machine learning models to make accurate predictions and investment recommendations. Furthermore, emerging trends in entrepreneurial finance, such as blockchain technology, coupled with the growing significance of sustainability, offer promising avenues for applying alternative data and machine learning applications in early-stage investment. These topics are explored in greater detail in the rest of the thesis.

Chapter 3

An investigation of sentiment analysis of information disclosure during Initial Coin Offering (ICO) on the token return

3.1 Introduction

The digital revolution has a remarkable impact on many aspects of life, and finance is no exception. In the last decades, firms and investors see their portfolio of opportunities widen and become more profitable on average. The intersection of finance and digital innovation makes available new financial tools such as equity or reward crowdfunding, peer-to-peer lending (Ahlers et al., 2015; Belleflamme et al., 2014; Wei Shi, 2018) and Initial Coin Offerings (ICOs) (Giudici and Adhami, 2019; Bellavitis et al., 2021; Fisch, 2019).

In this context, ICOs become tremendously important as a source of equity financing. An ICO can be defined as a sale of digital assets "tokens" (cryptocurrency), and since their first occurrence in 2013 (funding operation held by Mastercoin) ICOs prove to be extremely popular. The reasons are manifold (Andrieu and Sannajust, 2023), spanning from the hedging role of cryptocurrency with respect to the volatility of domestic currency and geopolitical risk (Momtaz, 2020), to the lack of trust and strict selection criteria for funding of the banking sector (Block et al., 2012). Pilkington (2018) adds their higher expected returns and lower transaction costs, whereas Adhami et al. (2018) also points out the increased liquidity of firms' investments due to the secondary market of tokens. Regarding their characteristics as a funding mechanism, ICOs prove to be suitable to finance projects characterised by a high uncertainty in their characteristics, potential and returns (Chen, 2019; Narayanan et al., 2016; Bellavitis et al., 2021), displaying positive effects of a built-in user base and network (Giudici and Rossi-Lamastra, 2018). The ICO market kept momentum in terms of volume and value in 2014, when Ethereum sold 31,000 bitcoins worth $ 18.3 million, (Rooney, 2018) and on, until 2018 when a decline (and recent recovery) took over as a result of the inability of regulation to match the growth of the market (Andrieu and Sannajust, 2023). Despite their popularity, fraudulent activities are more likely in an ICO environment than in alternative equity funding sources (Hornuf et al., 2022), and they are estimated as an extremely high percentage of their overall operations. For instance, Sapkota et al. (2020) estimate approximately 56% in number and 65% in value in the period 2014 – 2019; Liebau and Schueffel (2019) claims 80%

of ICOs are fraudulent in 2017. The reason why ICOs are particularly prone to fraud is twofold. First, their links with ventures, i.e., investment projects at an early stage, are highly innovative, and volatile. Then, ICOs embed into a fundraising activity a high-level technological component (e.g., using decentralized finance (DeFi)), which adds complexity and even more need for information with respect to a standard funding mechanism. Together with an inappropriate regulation, this leads to a perfect example of adverse selection and asymmetric information (Akerlof, 1970) as follows. On the selling side of the market (fundraiser) the lack of information requirement constitutes an incentive for bad projects to participate, whereas on the buying side, investors are driven away from the difficulty to fairly value projects that rely on an incomplete and potentially unreliable information set. The subsequent outcome is adverse selection, that is, (i) only lower quality projects are encouraged to apply for funding, (ii) a higher percentage of bad projects are funded with respect to alternative standard financial tools and on average at a higher cost. In such a landscape, given the increasing interest and potential for DeFi–based financial tools, the market itself tries to provide a solution.

ICO is operated in a lightly or unregulated market; hence, there is no requirement to audit or monitor the information disclosed to investors. Without accessing fundamental data to assess past performance in the form of financial and non-financial statements similar to public equities investment, many investors need to rely on token ratings to evaluate the proposed token generated by communities and experts. These ratings are widely used to screen potential scam tokens or speculate on which tokens successfully list on an exchange and generate a return. There are ongoing arguments regarding the reliability of the token ratings and the ability to reduce information asymmetry to assess the credibility of the token and future performance (Lee et al., 2021). However, many studies emphasise the weaknesses of the existing token rating system, underlying its non-transparent process and dependence on easily extractable information (Ofir and Sadeh, 2020; Florysiak and Schandlbauer, 2019). Thus, positive ratings may not necessarily guarantee potential financial returns. The factors that cause the discrepancy between token ratings and post-ICO financial performance are sufficiently studied. This leads to a research question aimed at understanding the factors that can explain the discrepancy between ICO ratings and token returns, and although this study is unable to find conclusive results regarding the specific factors that cause this discrepancy, its existence leads to suggestions for the token assessment process.

Assessing the true value and potential of Initial Coin Offerings (ICOs) is a significant challenge in the rapidly evolving blockchain landscape. While traditional token ratings and venture information from platforms, whitepapers, and social media offer some insights, their reliability in predicting long-term success, funding thresholds, or secondary market listings remains debated (Fisch, 2019; Campino et al., 2022; Adhami et al., 2018; Howell et al., 2019; Lyandres et al., 2022; Lipusch, 2018; Bourveau et al., 2022; Ante et al., 2018). A research gap persists in understanding how information disclosed during fundraising genuinely translates into long-term post-ICO financial performance. With the latest advancements in artificial intelligence, there is an opportunity to integrate natural language processing (NLP) techniques for extracting quantitative information, such as sentiment analysis, from the text presented on whitepapers (structured by IPO prospectus categories) and social media platforms like Twitter, which is highly used by ventures. Hence, the study extends Bourveau et al. (2022)'s model by analysing whitepaper sentiment, integrating it with social media's emotional cues for 6-months return prediction. These refined signals are then integrated with previously studied factors to construct a novel ICO index, enhancing transparency beyond traditional ratings.

The study then deploys machine learning (ML) to forecast post-ICO token returns. Although some studies explore decision models to predict fundraising success (Lahajnar and Rozanec, 2018; Deng et al., 2018), precisely predicting post-ICO returns and minimising human bias inherent in community-driven token ratings remains a challenge. This research makes several contributions: it empirically confirms the discrepancy between ICO ratings and long-term token returns, underscoring the need for more robust token assessment. It highlights the significant role of social media sentiment in predicting and explaining token returns. Importantly, it demonstrates the impracticality of relying on whitepaper sentiment grouped by IPO prospectus for prediction, despite its initial conceptual appeal. By creating a novel ICO index and leveraging machine learning, this research provides an accurate and transparent framework for evaluating token value, moving beyond the limitations of traditional methods and significantly reducing information asymmetries in token financing.

The chapter is organised as follows: Section 2 outlines the background of token financing, and it lays down the hypotheses for the analysis. Section 3 introduces the dataset, and Section 4 highlights the methodologies for this paper, including NLP techniques and regression model. Section 5 presents and discusses the results of the statistical analysis, followed by classification results in Section 6. Lastly, Section 7 concludes the paper and outlines future work.

3.2 Literature review and hypotheses formulation

Blockchain technology, disrupting the path dependence of centralised networks such as the Internet and social media platforms, challenges the persistence of technology use following past events (Arthur, 1989; Schilling, 2002). The proposal of Bitcoin in 2008, leveraging the decentralisation and peer-to-peer properties of blockchain, creates significant value in financial industries, reaching a peak market capitalisation of one trillion in 2018 (CoinMarketCap, 2022). The financial market embraces this innovation by introducing new assets in the form of cryptocurrency, serving as a digital medium of exchange. Moreover, entrepreneurs and ventures capitalise on this innovation through token financing methods, notably Initial Coin Offering (ICO), with the potential to become a mainstream financing avenue.

3.2.1 ICO Process

In contrast to traditional equity financing, ICO raise funds by issuing cryptographically secure tokens to investors. These tokens, with characteristics specified by issuers, can be categorised as cryptocurrency, utility tokens, and security tokens. Cryptocurrency serves as a medium of exchange for goods and services, utility tokens grant access to the blockchain networks, and security tokens represent ownership of assets (Howell et al., 2019). The decentralised feature of ICO is associated with lower transaction fees and token price appreciation, driven by the network effect as more counterparties become available for transactions, leading to increased demand and higher token prices (Cong et al., 2020).

Another key distinguishing characteristic is a lightly regulated nature of the token financing market compared to traditional equity financings like Initial Public Offering or IPO (initial public offerings), transactions must follow regulatory jurisdictions, necessitating the disclosure of financial and non-financial information to prospective investors in compliance with rules. Consequently, IPO involves high costs, lengthy processes, administrative and legal requirements, and multiple agents. However, the true cost of token financing is still unclear (Andrés et al., 2022). To address information disclosure gaps, ventures opt for alternative approaches to attract potential investors in exchange for tokens. Entrepreneurs can publish relevant information on websites and social media platforms such as Twitter, Reddit, and Medium, facilitating interaction between token issuers and communities. Additionally, ventures can choose to release a whitepaper, a voluntary disclosure document offering project overviews, teams' details, and financial strategies. Alternatively, startups can list on ICO database platforms such as ICObench, ICOdrop, and ICOdata at no cost. These platforms employ rating mechanisms based on voluntary information disclosure and expert's evaluations. The availability of information implies the assumption that it has the potential to mitigate information asymmetry in token financing and reduce the analytical risk of choosing the wrong ICO due to a lack of information (Karpenko et al., 2021; Cong et al., 2020; Campino et al., 2022; Burns and Moro, 2018; Fisch, 2019; Ofir and Sadeh, 2020).

The ongoing demand for investing in ICO tokens can be attributed to behavioural finance biases, including overconfidence in herd behaviour and the fear of missing out (FOMO) (Liu, 2019). From a pricing perspective, the ICO mechanism encourages buyer competition for tokens due to their scarcity. Furthermore, investors can value or price these tokens without shareholder rights compared with traditional equity (Catalini and Gans, 2018). Catalini and Gans (2018) argue that the value of a token depends on a single period of demand, and the absence of a commitment to hold the token for an extended period allows for flexibility in future funding, unlike illiquid equity financing for early-stage companies, which often occurs in multiple rounds. Despite the increased flexibility in financing, token returns witness a decline overtime. This phenomenon can be attributed to investors learning from their experiences, leading to a reduction in underpricing compared to the first day of trading, which typically involves greater uncertainty and urgency in token distribution (Benedetti and Kostovetsky, 2021). Overall, the potential mainstreaming of token financing for early-stage companies prompts researchers to examine the determinants influencing ICO success and the potential value of post-ICO.

3.2.2 Determinants of ICO success and return

The ICO literature defines drivers of the success of ICO and analysed various sources of fundamental data, including ICO characteristics, social media and whitepaper.

ICO characteristics, available on the token issuer's website and ICO databases, often include project details, founder information, and fundraising campaigns. Many studies examine the relationship between these ICO characteristics and the success of token financing campaigns, particularly in terms of the likelihood of reaching funding goals. Campino et al. (2022); Deng et al. (2018); Amsden and Schweizer (2018); Howell et al. (2019) identify human capital features like team size, LinkedIn network connection, advisors availability, and founder quality as positively correlated with token funding success. However, Deng et al. (2018); Howell et al. (2019) argue that possessing these characteristics does not necessarily guarantee the successful delivery of products and services to the market. Campino et al. (2022) find that a high token price with bonuses is perceived as a quality signal to investors, while cheap tokens are considered potential scams. Additionally, the level of token retention held by entrepreneurs and the team, indicating a commitment to establishing the company, is positively correlated with ICO success (Amsden and Schweizer, 2018; Davydiuk et al., 2023). Pre-selling of tokens before ICO, allowing investors to purchase at a discounted price, may occur; however, this practice can lead to flipping and pump and dump schemes, and market manipulation (Andrés et al., 2022). Overall, while previous studies focus on the success of meeting the minimum funding threshold as the dependent variable, there exists a research gap in further analysing the effects of these signals on token returns. Another signal that can be considered is the ICO analysis ratings, which rely on easy-to-extract characteristics shared by ICO ventures.

Lee et al. (2021) investigate ICO aggregated ratings, a collective opinion from communities and experts in token financing, and found associations with fundraising success and long-run token return. Bourveau et al. (2022) note that ICO rating platforms and experts serve as market-based information intermediaries with strong reputations for accurately rating ICO ventures. However, this rating information may become commercialised, potentially resulting in market inefficiency, as the rating might not adequately consider the technical aspects of the project (Ofir and Sadeh, 2020). This aligns with Florysiak and Schandlbauer (2019)'s findings that the evaluation process relies on easy-to-extract information, such as the team size, the existence of KYC, the number of active social media, and the presence of the whitepaper, without necessarily understanding its content. Thus, the effectiveness of ICO ratings in reducing information asymmetry and the quality of the assessment method is open to question. Despite these findings, Lee et al. (2021) report contradicting results, suggesting that the aggregated rating is associated with the token return. However, this raises the question of whether a strongly positive rating necessarily guarantees a positive token return and to this aim, this study proposes the following hypothesis:

> H_1: There is a negative relationship between the rating assigned to an ICO by the platform and its six-month return.

To address the potential discrepancy between ratings and token return as a proxy of its fair value, the study aims to extend the model to understand the determinants of token return and consider other sources of information, such as the whitepaper relying on information published by token issuers rather than a collective opinion by communities.

As previously mentioned, the rating often lacks consideration of the technical aspects of the project, which are typically detailed in the whitepaper. This observation opens another research opportunity, as many studies investigate the role of whitepapers. Whitepapers serve as documents prepared by token issuers to outline products and services, similar to traditional business plans, providing information to their potential investors (Lipusch, 2018). Cong et al. (2020); Campino et al. (2022) conclude that whitepapers play a crucial role in reducing information asymmetry and informing investors. Several studies examine the impact of whitepapers on the amount of capital raised during an ICO campaign. Ante et al. (2018); Bourveau et al. (2022) find that the existence of a whitepaper strongly influences the amount of capital raised. Fisch (2019) delve into the role of technical pages, which focus on technology, complement and architecture design, and find that these elements influence the amount raised as it attracts investors. Other elements such as the number of pages and citations do not significantly impact ICO success, suggesting that investors may consider the mere existence of a whitepaper as a signal rather than delving into the document's details (Ante et al., 2018). This conclusion aligns with the findings of Liu (2019); Adhami et al. (2018); Florysiak and Schandlbauer (2019), who report that whitepapers do not significantly affect funding success and token investment returns. The lightly regulated nature of the ICO market contributes to issues related to whitepapers. There are no formal requirements for the content included in whitepapers, leading to challenges such as the absence of certification, standard format or authority to audit the information presented (Ante et al., 2018). Consequently, whitepapers may be less valued by ICO ventures.

Despite challenges in analysing whitepapers, quantitative attributes such as the sentiment can be extracted from the text. For example, Florysiak and Schandlbauer (2019) study the effect of sentiment analysis on 22-day and 66-day token return and volume. However, their studies focus on overall content that can be subdivided into sections and can be further investigated with the long-term post-ICO return. The paper further examines the relationship between the sentiment of the whitepaper and post-ICO return over an extended period of up to six months, studying potential effects that are not previously investigated. In addition, the study extends to extract the sentiment of the whitepaper based on each aspect following (Bourveau et al., 2022)'s methodology, which previously examines the availability of the information presented in the whitepaper following the IPO prospectus. To investigate the potential role of the whitepaper in driving the returns of the ICOs after six months, the research formulates the following hypotheses regarding the different categories of its content:

> H_{2a}: There is a positive relationship between the technical characteristics of the project (henceforth "Technology") as described in the whitepaper and the six-month return of the ICOs.
>
> H_{2b}: There is a positive relationship between the characteristics of the team (henceforth "Team") and the six-month return of the ICOs.
>
> H_{2c}: There is a positive relationship between the economic and financial characteristics of the project (henceforth "Market") and the return of the ICOs after six months.

The social interaction of agents in the environment can influence decision-making and economic outcomes through the transmission of biases (Han et al., 2022). This phenomenon can be seen in the social media set up by the ICO venture to communicate and share information about products and services, overcoming information asymmetries. The management and development team engage with the community, including potential investors and end users, via well-known social platforms such as Twitter, and Facebook, or messaging apps like Discord, Slack, and Telegram (Lipusch, 2018; Bourveau et al., 2022). Bourveau et al. (2022) identify Twitter as the most popular social media tool among ICO ventures, with 97% having a Twitter presence. Fisch (2019); Bourveau et al. (2022); Lyandres et al. (2022); Campino et al. (2022); Burns and Moro (2018); Ofir and Sadeh (2020); Ante et al. (2018) examine Twitter activities, analysing metrics like the number of followers and the number of tweets posted by token issuers. These studies found positive correlations with ICO funding amount raised, emphasising the role of tweets in reducing information asymmetry. However, Lyandres et al. (2022) report contradictory results, finding that the number of tweets is not statistically significant for funding success. Stanley (2019) explore the sentiment of tweets and the level of social media activity during the pre-ICO and found no correlation between the return on investment. The research findings suggest that social media activity posted before starting an ICO campaign may not necessarily impact the token value. Despite this, there is little research evaluating the tweet sentiment during ICO fundraising and its impact on long-term post-ICO returns. This research gap raises another question for an investigation into the sentiment of tweets and its relationship with long-term post-ICO returns. To investigate the role of tweets in driving the ICO returns, the study formulate the following hypothesis:

> H_3: There is a positive and statistically significant correlation between the sentiment about the project as proxied by the tweets and the six-month return of the ICOs

The absence of regulation in the ICO space leads researchers to explore ways to assess the trustworthiness of tokens as investment opportunities. Deng et al. (2018) construct token valuation measurement, highlighting the influence of technological development and user base on token value. Deng et al. (2018); Bourveau et al. (2022) also create a disclosure index and found a positive correlation between the funding success of unrated ICOs and token returns. While many researchers delve into the factors contributing to ICO success and the amount raised, there is limited exploration of applying machine learning to predict the post-ICO returns. Such models are well-suited for handling large volumes and diverse types of data, and they can generate predictive insights without direct human intervention. This study proposes a model that aims to predict post-ICO performance based on information disclosed during the fundraising period, including sentiment extracted from text data. This approach leverages the capabilities of machine learning to enhance understanding and prediction in the ICO context.

3.3 Data description

Initial Coin Offerings

ICObench is one of the popular rating platforms that provide ICO data, and its ratings are extensively utilised and studied in numerous previous works (Lee et al., 2021; Lyandres et al., 2022; Bourveau et al., 2022). This 5-star rating platform employs its ICO assessment algorithm that incorporates independent experts who voluntarily rate the token for the token financing community (ICObench, 2017). The score weight is distributed according to the number of experts and their past contributions as active members of the ICObench community. The ratings are assigned based on the information disclosed during the ICO without incorporating the market data once the venture successfully raises funding.

Since 2015, when the first ICO was advertised on the ICObench platform, a total of 5,723 ICOs[1] are listed on the website. The study aims to investigate the financial performance of post-ICO; thus, only venture campaigns that are successfully listed on the secondary market or crypto exchange are analysed. Some ICOs are removed due to incomplete information on crucial characteristics, such as a lack of Twitter presence and the unavailability of token prices, which is elaborated upon later. Consequently, there are 391 tokens available for analysis, and A includes the list of variables used in this study.

Whitepaper

A whitepaper serves as a document outlining key details of the blockchain project, such as the algorithms, the product development roadmap, the management and development team, revenue models, and tokenomics. Typically, a whitepaper is in PDF format and is publicly accessible, being uploaded on ventures' websites, ICO rating platforms and cryptocurrency exchanges. However, their provision is voluntary, leading to variations in disclosure requirements among ICO ventures. Also, there are technical issues while extracting the text, such as the embedded information as an

[1] The data is collected on 11 February 2022.

Table 3.1: The description of whitepaper aspects following the Bourveau et al. (2022) technique, which studies the existence of information in each IPO prospectus.

Whitepaper Aspects	Description
Product Description	Description of venture primary business's purpose and outline how the blockchain product and services would solve a particular problem.
Technical	Description of algorithm and architecture of blockchain application and tokens.
Team	Description of human capital involved in the project: management team, developer team, and advisors. The information includes their identity and professional background.
Product Roadmap	Description of products and features to be developed and delivered in each milestone.
Finance	Description of token sale plan, which indicates the token allocation, incentives for token sales and how the fund will be used for the expense.
Business Landscape	Description of market and industry research, market position, product matrix and competitor analysis.
Risk	Description of risks involved in investment and risks related to the products and services such as technology and regulatory risks

image. Any records with missing and non-English text are removed from the analysis. As a result, only 161 out of 391 ICO ventures have whitepapers available.

This study extends Bourveau et al. (2022)'s model by analysing the sentiment of the whitepaper by IPO prospectus rather than the binary variable of whitepaper availability. Table 3.1 shows the whitepaper categories, including additional aspects such as the business landscape and risk factors.

Tweets

Twitter is a social media platform widely used by numerous ICO ventures for communication, community engagement, and attracting potential investors (Bourveau et al., 2022). The tweet message length is shorter than other social media with a maximum length of 280 characters. Tweets are collected via Twitter API and twarc, a Python package (Summers et al., 2014). Only the tweets posted by the venture and during the token funding are collected for sentiment analysis. This study is designed to examine the impact of information disseminated by ICO ventures. Accordingly, the research focus is solely on the analysis of tweets originating from these ventures, excluding those from other Tweets users expressing their opinions on the token through hashtags or mentions of the token issuer's account. The rationale behind omitting tweets from other Twitter users is driven by the commitment to following privacy policy during the data collection phase. The dataset comprises a total of 17,520 tweets emanating from ICO ventures, representing 391 distinct tokens.

Token Price

The final dataset employed in this study is the token price data. As each token can be traded on single or multiple exchanges, establishing standard prices across exchanges presents a challenge, given variations in governance that can range from decentralised to centralised exchanges. Fortunately, Coingecko (Coingecko, 2022) addresses this challenge by aggregating prices through a value-weighted average function that considers exchange pairings. The platform also consolidates token trading volumes across exchanges and offers market capitalisation in USD - the ratio between the current crypto-asset price and the available supply. For the purposes of this analysis, volumes and market capitalisation are beyond the scope. The study aims to assess changes in token prices traded on an exchange for at least six months after the end of ICO. The price data is used to compute a token return using the following function:

$$log\_return = \log \frac{price_j}{price_i} \tag{3.1}$$

where $price_j$ is the token price after six months, and $price_i$ is the token price on the first day of trading. The result of Formula 3.1 provides the rate of return, normalised to share the same scalar even when the price values of each token vary. This log return serves as the dependent variable in the regression model and the target variable used in the classification model. However, it is important to note that not all tokens listed on ICObench have price data available on Coingecko. Some ICOs may fail to raise sufficient funds and reach soft capitalisation, or they might be later delisted from the exchange.

3.3.1 Statistics summary result

Out of the 5,723 ICOs listed on ICObench, 391 tokens meet the criteria and are analysed in this study. These ICOs have token price data traded on secondary markets or crypto exchanges for at least six months and have Twitter accounts. The majority of ICOs (47.8%) commenced in 2018. These blockchain applications aim to address challenges across industries, including banking, healthcare, and entertainment, with 20.9% identifying as platforms, and 13.8% as cryptocurrencies. Over 80% of ICOs develop their products and services on Ethereum, a smart contract blockchain adhering to the ERC (Ethereum Request for Comments) standard.

The applications are developed and operate in various geolocations; the majority of ICOs are established in Singapore (16.8%), followed by the United States at 15.0%, and the United Kingdom at 6.72%. Nevertheless, some ventures also initiate operations in countries with unfavourable regulations on token fundraising and cryptocurrency, such as China (2.33%). Additionally, ventures may encounter operational restrictions in certain countries. The average number of restricted countries, excluding the country that completely bans cryptocurrency, is one. Examples of restricted countries included the United States, South Korea and Japan.

Based on information published on ICObench, other aspects of the ICO include the venture team, financial information, and tokens issue during the fundraising period. The average team size is approximately 15 members, with 68.5% and 50.9% of ICOs having a Chief Executive Officer (CEO) and a Chief Technology Officer (CTO), respectively. The CEO and CTO are the key players in the development of blockchain applications, overseeing and monitoring the project and organisation to deliver the promised products to potential investors. While the CEO's engagement in a prior blockchain project signals positively for ICO ventures, yet only 1.80% of CEOs have directly comparable previous experiences.

As previously mentioned, every ICO of the dataset used in the study has a social media presence on Twitter. Additionally, the ICO ventures utilise other communication channels such as Facebook, Discord, Telegram, Slack, and Reddit, averaging around seven platforms per project. Another method to gain investor confidence is through the KYC (Know-Your-Customer) process, ensuring the verification of ICO venture team members. Alternatively, ICOs can implement a whitelist, requiring investors interested in token sales to undergo KYC procedures (Taçoğlu, n.d.). More than 58.3% of ventures have a whitelist or KYC process available for token selling.

Other approaches to token selling and distribution include Initial Exchange Offerings or IEO and pre-selling of tokens. Approximately 10.7% of the tokens are listed through IEO. The objectives of both ICO and IEO align in securing capital for blockchain projects. However, the token sale process of an IEO occurs on cryptocurrency exchanges such as Binance Launchpad and OK Jump start Launchpad (Binance Academy, 2020). Consequently, investors tend to have more confidence in purchasing assets through these exchanges compared to venture websites or token platforms. Over 47.1% of ventures allow the investor to participate in pre-selling or pre-ICO, often offering incentives or bonuses during this period to attract investors. Regarding processing times, the fundraising process takes an average of 51 days to complete, followed by ventures spending approximately 119 days, or almost four months, to list these tokens on exchanges. As previously explained, due to limited regulation and monetary policy controlling liquidity and the variety of exchanges available for token trading, ventures can list the tokens on numerous exchanges. Hence, the average number of markets per project is six exchanges.

The information for each ICO project is not consistently available on ICObernch. Out of 391 ICOs, 128 share details regarding the soft capitalisation or the minimum fundraising target for the project. The figure usually depends on the project's needs and plans for developing and improving products and services or expanding into the market. The average log value in USD is 14.9. Additionally, 302 ventures outline the distribution of tokens sold during the ICO, with an average distribution of 46.1%. This indicates that 46.1% of tokens are available for investors to purchase during the ICO, while the rest may be allocated to founders and team, advisors and company reserves. Furthermore, 310 token issuers disclose the number of tokens for sale to investors, with the mean of its log value being 19.5. Table 3.2 provides a summary of the statistics for each variable used in the study.

Table 3.2: Summary Statistical Table of ICO variables captured from ICObench, Whitepaper, Twitter, and Coingecko.

Name	Count	Mean	Std	Q25	Q50	Q75
Panel A: ICO Campaign						
Soft Cap (log)	128	14.917	1.631	13.952	14.914	15.676
Whitelist/KYC	391	0.583	0.494	0.000	1.000	1.000
Pre-ICO	391	0.471	0.500	0.000	0.000	1.000
IEO	391	0.107	0.310	0.000	0.000	0.000
Bonus	391	0.396	0.490	0.000	0.000	1.000
ICO Duration	391	51.000	92.460	10.000	29.000	55.000
Token Listing Duration	391	119.248	198.976	20.000	55.000	116.500
Panel B: Token						
Token for Sale (log)	310	19.475	2.290	18.118	19.382	20.352
% Token Sale	302	46.057	22.566	30.000	50.000	60.000
ETH-based	391	0.831	0.375	1.000	1.000	1.000
Panel C: Team						
Team Size	391	15.898	9.577	10.000	15.000	19.000
CEO	391	0.685	0.465	0.000	1.000	1.000
CTO	391	0.509	0.501	0.000	1.000	1.000
CEO Prev Experience	391	0.018	0.133	0.000	0.000	0.000
Panel D: Whitepaper						
Whitepaper Disclosure	391	0.412	0.493	0.000	0.000	1.000
Problem Description Disclosure	154	0.935	0.247	1.000	1.000	1.000
Technical Disclosure	158	0.943	0.233	1.000	1.000	1.000
Roadmap Disclosure	97	0.959	0.200	1.000	1.000	1.000
Team Disclosure	89	1.000	0.000	1.000	1.000	1.000
Financial Disclosure	102	0.990	0.099	1.000	1.000	1.000
Business Landscape Disclosure	74	0.959	0.199	1.000	1.000	1.000
Risk Disclosure	33	0.394	0.496	0.000	0.000	1.000
Panel E: Social Media						
Social Media	391	7.197	1.641	6.000	8.000	8.000
Twitter Activity	391	44.808	93.191	7.000	19.000	43.500
% Positive Tweets	391	91.600	12.076	88.118	95.141	100.000
Panel F: Market						
Market Size	391	11.307	20.733	2.000	4.000	8.000
Restricted Area	391	1.852	5.304	0.000	0.000	1.000
Panel G: Rating						
ICO Rating	391	3.419	0.653	3.000	3.400	3.900
Panel H: Token Price and Return						
Token Price Day 1	391	1.989	21.397	0.011	0.055	0.305
Token Price Day 30	391	1.889	20.694	0.009	0.051	0.301
Token Price Day 90	391	1.838	16.987	0.006	0.037	0.245
Token Price Day 180	391	2.763	34.825	0.004	0.022	0.192
First day Token Return (log)	391	-0.014	0.271	-0.093	-0.010	0.071
180 days Token Return (log)	391	-0.742	1.765	-1.816	-0.922	0.288

3.4 Methodology

3.4.1 Natural language processing and sentiment analysis

This study examines two text datasets: whitepaper and tweets. However, these text data can not be directly input into statistical and machine learning models without preprocessing into numerical representations. The recommended approach involves removing noises, such as stopwords and punctuations, followed by tokenisation into word tokens and transformation into their base forms using lemmatisation. Tweet messages may contain emojis and hashtags referring to specific topics, URLs and mentions of other users. These elements do not contribute linguistic value to understanding the sentiment, whether positive or negative and are therefore removed during text preprocessing. Various natural language processing (NLP) techniques can process text vectors to enable sentiment analysis, a binary classification task that categorises text into positive or negative sentiment. However, sentiment analysis approaches typically require a training dataset, which is impractical for the target domain to gather. To address this limitation, this study adopts the Valence Aware Dictionary of Sentiments (VADER) (Hutto and Gilbert, 2014).

VADER is a lexicon-based sentiment analysis tool that bypasses the need for a training dataset, making it suitable for the study. VADER is designed and developed to offer several advantages of rule-based modelling. It provides a generalisation of social media sentiment that can be applied across multiple domains without a training dataset. Furthermore, the performance of the model is claimed to be efficient without compensating speed, making it suitable for processing online streaming data with high velocity. The VADER process begins by constructing a list of lexical features and assigning sentiment intensity to each feature on a scale from strongly negative to extremely positive. The list considers common expressions in microblogs, including emoticons and sentiment-related acronyms. Additionally, five generalisable heuristic rules are developed and integrated with the list of lexical features. These rules are based on grammatical and syntactic components, linking attributes to sentiment intensity and accounting for word order – a sensitivity that the bag-of-words model may not capture.

The VADER model produces a compound score, which can be interpreted for binary classification into positive and negative sentiments. If the compound score is positive, the text is labelled it as positive sentiment; otherwise, it is considered negative. Previous studies show that VADER outperforms other lexicon-based models, including OpinionFinder (Wilson et al., 2005) and AFINN (Nielsen, 2011), in various domains, including the financial domain (Sazzed and Jayarathna, 2021; Sohangir et al., 2018). VADER is also successfully applied in the cryptocurrency domain; for example, Kraaijeveld and De Smedt (2020) use VADER to capture Twitter sentiment and forecast cryptocurrency prices of Bitcoin, Bitcoin Cash and Litecoin.

This study applies VADER to assess the polarity and intensity of emotion expressed by ICO ventures in tweets. The performance of VADER is analysed, considering that its application is not explicitly outlined for non-social media text, such as whitepapers, which may contain linguistic expressions different from those found in social media text.

3.4.2 Correlation

The study conducts correlation and regression analyses to understand the relationship between the independent variable representing ICO characteristics and information disclosure and the dependent variable of 6-month token investment return. This study incorporates three types of variables: numerical, categorical, and ordinal values in Boolean. Each type of variable requires a different correlation analysis —Pearson correlation for numerical variables and Point Biserial correlation for categorical and ordinal variables. The Pearson correlation involves computing the covariance of two continuous variables (x and y) divided by the product of their respective standard deviation. Additionally, the Point Biserial coefficient, which shares the same computation as the Pearson correlation, assesses the degree of association between categorical and continuous variables.

$$log\_return_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon \tag{3.2}$$

X_i = independent variable (see table 3.2)
ϵ = residuals $\sim N(0,\sigma^2)$

However, while correlation measures the strength of the relationship, regression analysis is employed to quantify the association between the two variables. Multiple linear regression (Formula 3.2) estimates the coefficient, which explains the relationship and the effect of the independent variable on token return, utilising Ordinary Least Square (OLS) to find coefficients that minimise the error in predicting the dependent variable.

3.4.3 ICO Index

The study aims to explore another dimension by constructing a novel ICO index that considers characteristics and information disclosure during token financing. This approach combines signals used by ICObench and those examined in previous literature, with the goal of enhancing transparency compared to existing ICO ratings (ICObench, 2017; Deng et al., 2018; Lahajnar and Rozanec, 2018; Bourveau et al., 2022). Unlike the disclosure index created by Bourveau et al. (2022), which is an arithmetic sum of voluntary disclosures from whitepapers, source code, and social media used to investigate the positive correlation with ICO fundraising success (investment decision and quantity), this novel index focuses specifically on the relationship between its scores and potential token returns.

The index comprises categories such as Teams, outlining team size and key players like the CEO and CTO. Additionally, it assesses technology and campaign strategies, including the availability of IEO, pre-selling, bonuses and the disclosure of technology aspects in the whitepapers. The third factor encompasses the market, considering the size of the markets where the startup operates and any restricted areas of operation. Lastly, the index includes the ratio of positive messages found on Twitter.

To construct this index, studies streamline categories by selecting variables that pass multicollinearity tests and possess complete information. The ICO index is then constructed as a linear combination, derived from the sum of its variables, with each contributing equally, as detailed in Table 3.3. Continuous variables, such as

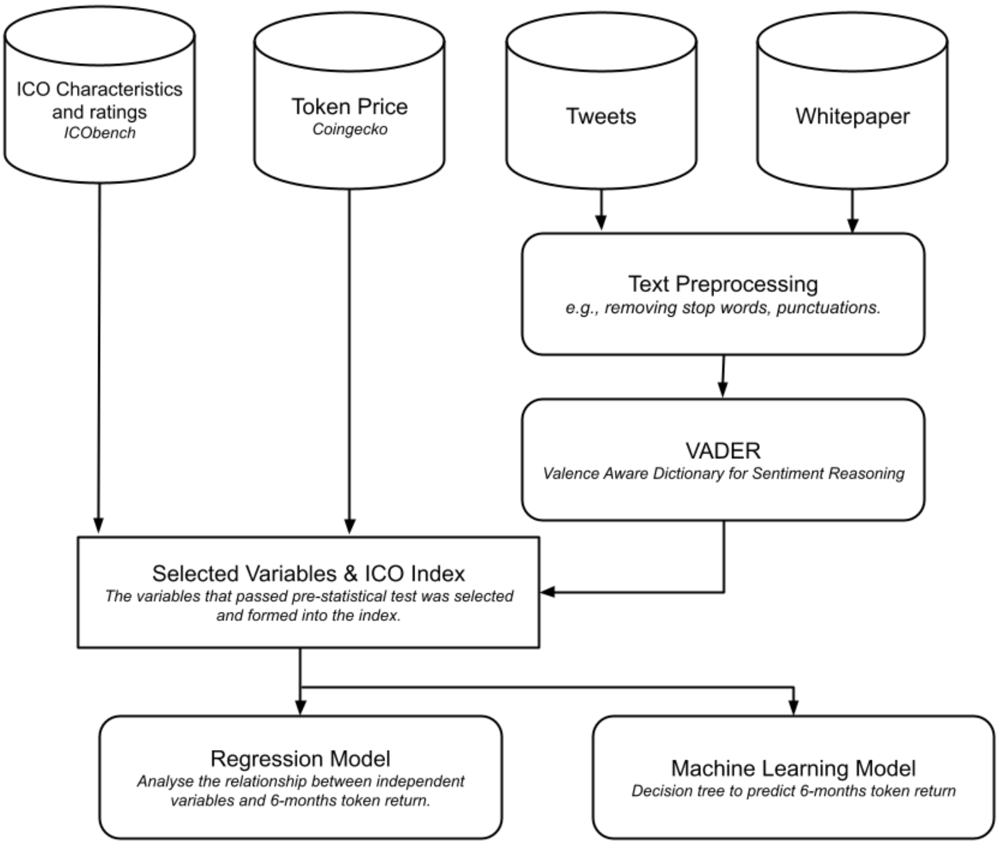
ICO Characteristics and ratings
ICObench
Token Price
Coingecko
Tweets
Whitepaper
Text Preprocessing
e.g., removing stop words, punctuations.
VADER
Valence Aware Dictionary for Sentiment Reasoning
Selected Variables & ICO Index
The variables that passed pre-statistical test was selected and formed into the index.
Regression Model
Analyse the relationship between independent variables and 6-months token return.
Machine Learning Model
Decision tree to predict 6-months token return

Figure 3.1: A diagram shows the sources of data collection and the methodology used in this study. The variables collected from various data sources are passed into a pre-statistical test before being selected to test in the regression model and machine learning and to construct the ICO index.

ICO Duration and Token Listing Duration, are transformed into categorical variables based on their median benchmarks. This table also outlines the specific categories of variables, including Team, Technology, Campaign, Market, and Social Media. Subsequently, an in-depth exploration of this novel ICO index is conducted to analyse its impact on the token return through regression analysis. This investigation also evaluates whether the inclusion of the ICO index improves the robustness and prediction accuracy of token return, utilising machine learning models.

The data collection process and methodologies are illustrated in Figure 3.1, with further details on the machine learning model provided in Section 6.

Table 3.3: A novel score category that aggregates the ICO variables used in the study. Some variables are transformed from continuous values into dummy values whether the value higher or lower than the threshold.

Score Category	Name	Description
Team	Team Size	A dummy variable of one of the team sizes is greater than the median; otherwise, zero.
	CEO	An indication of whether the venture has a Chief Executive Officer (CEO).
	CTO	An indication of whether the venture has a Chief Technology Officer (CTO).
	CEO Prev Experience	An indication of whether a CEO has prior experience in blockchain projects.
Technology and Campaign	Whitelist/KYC	An indication of whether the Know-Your-Customer (KYC) and whitelist are performed prior to purchasing the token.
	ETH-based	An indication of whether the application is operated on Ethereum blockchain.
	Pre-ICO	An indication of whether the pre-ICO or pre-selling is available.
	IEO	An indication of whether the token sale is handled and vetted by an exchange, i.e., Initial exchange offering (IEO).
	Bonus	An indication of whether the bonus is available for investors.
	ICO Duration	A dummy variable of one if the number of days for a token to be available on exchanges is less than the median; otherwise, zero
	Token Listing Duration	A dummy variable of one if the number of days taken during the ICO process is less than the median; otherwise, zero
	Whitepaper Disclosure	An indication of whether the venture has a whitepaper.
Market	Market Size	A dummy variable of one if the number of markets or exchanges that sell the token is greater than the median; otherwise, zero.
	Restricted Area	A dummy variable of one if the number of countries restricted for the project to operate is less than the mean; otherwise, zero.
Social Media	% Positive Tweets	A ratio of positive sentiments of tweets posted by the venture.

3.5 Result

3.5.1 Sentiment analysis result

Prior to a detailed explanation of the results addressing each research question, this study presents an overview of the sentiment analysis results achieved through the application of VADER techniques on both tweet and whitepaper text.

As previously explained, VADER is a lexicon and rule-based sentiment analysis tool that can give the sentiment score of each text. The compound score generated by VADER sums neutral, positive and negative sentiments and normalises within the range $[-1, 1]$, where -1 indicates extremely negative sentiment and 1 for strongly positive sentiment. The non-negative compound score is interpreted as positive sentiment and negative otherwise. This study applies VADER on two text datasets: whitepaper and tweet messages.

For tweets, 17,520 messages representing 391 ICO ventures are analysed, and 92.0% of tweets are classified as positive, while the rest conveyed a negative sentiment. Moreover, each ICO had an average ratio of the positive sentiment of overall tweets at 91.6%. The result indicates a positive sentiment toward tweets posted by the token issuer. These sentiment ratios can be further used in statistical analysis to study their relationship with token returns. This research specifically focuses on analysing the emotion based on the text shared by the ICO ventures.

After applying VADER and capturing the sentiment of the whitepaper, which has different linguistic characteristics from tweets, it is found that only 161 out of 391 ventures provide the whitepaper, and each whitepaper does not necessarily cover all information aspects as previously described in Table 3.1. Table 3.4 shows that only 158 of 391 tokens had a whitepaper that discloses technical information about blockchain technology and architecture design. Of these, 154 ventures describe the problem statement and how blockchain technology would address industry issues. About 63.3% outline the token issuing and selling process and how the financial budget would be distributed during development. However, only 20.5% of the whitepapers outline investment-related risks, including technological and operational risks. This specific section of the whitepaper yields the highest negative sentiment ratio at 60%.

Table 3.4: The summary of statistics of a whitepaper published by token issuers with the number of whitepapers that have information aspects outlined in table 3.1, number of positive sentiment and negative sentiment, and the top ten words found in each whitepaper aspect.

Whitepaper Aspects	Observations	Positive Sentiment	Negative Sentiment	Top ten Frequency Words
Problem Description	154	144	10	blockchain, data, user, use, technolog,platform, market, network, token, develop
Technical	158	149	9	data, use, user, transact, network,blockchain, token, node, contract, servic
Team	89	89	0	develop, token, platform, launch,transact, payment, market, user, releas, exchang
Roadmap	97	93	4	develop, manag, blockchain, team, busi,technolog, experi, market, year, compani
Finance	102	101	1	token, sale, develop, use, user,platform, market, distribut, fund, particip
Business Landscape	74	71	3	market, user, platform, use, game,data, develop, token, busi, servic
Risk	33	13	20	token, may, risk, platform, compani,purchas, includ, develop, use, attack

The impact of sentiment extracted from both tweets and the whitepaper is subject to further investigation through regression modelling.

3.5.2 Discrepancy between rating and return

To validate the existence of a discrepancy between ratings and token returns, an analysis of a dataset of 391 tokens is conducted. The results reveal that the assumption linking positive (negative) token ratings to positive (negative) financial returns is not necessarily valid (Table 3.5). This assumption is previously explored in the literature. On the first day of listing on exchanges, 42.7% of tokens from ICOs rated as good (with a rating greater than or equal to 2.5) generate negative returns. Conversely, 53.33% of ICOs considered as poor, with ratings below the benchmark, yield positive returns. This disagreement between rating and return persists consistently across various timeframes, extending up to six months. The number of ICOs with negative returns increases from 14 to 20 tokens six months after the initial trade date when the tokens became available on crypto exchanges. The results suggest that a good (bad) rating does not necessarily guarantee future positive (negative) returns. These findings underscore the limitations of ICO ratings, signalling that investors should not rely solely on them as a singular indicator of potential token returns after fundraising. This is extremely important also at a regulatory level suggesting a twofold approach. On one hand, in their capacity as drivers for investment decision-making, ICO ratings should be more transparent in their components and technical aspects. On the other hand, their misleading nature is an additional call for greater information disclosure asked of the issuers.

The next section explores the relationship between information disclosure during the ICO and token returns, aiming to provide insights into this observed discrepancy, which is further detailed in the regression analysis.

Table 3.5: A number of tokens with a rating given on the ICObench website and classified by the token return of 1 and 180 days after the token is listed on the exchange.

1 day	Positive Return	Negative Return
Rating ≥ 2.5	154	207
Rating < 2.5	16	14
180 days	Positive Return	Negative Return
Rating ≥ 2.5	108	253
Rating < 2.5	10	20

3.5.3 Regression analysis result

In this study, the OLS regression is employed to examine the relationship between ICO information disclosure and six-month token return on the exchange. The dataset consists of two groups of observations: the first dataset includes 391 ICOs, and the second dataset comprises 263 tokens that are identified to have a discrepancy between the rating and the token return after six months. This discrepancy attribute implies that the token receives ratings higher (less) than 2.5 and experiences a negative (positive) return.

Pre-statistical testing identifies multicollinearity, indicating a high correlation among independent variables. Additionally, the study incorporates token type and country as control variables in the regression model. The token type is chosen to capture nuanced characteristics of different blockchain platforms, while the country variable controls for geographical disparities. These controls are implemented to standardise variations across diverse blockchain environments and isolate the true relationship between independent variables and six-month token returns. The strategic inclusion of these controls enhances the precision and interpretability of the findings, ensuring a more accurate estimation of relationships within the dynamic landscape of blockchain ventures. Consequently, only the variables listed in Table 3.6 are included in the Ordinary Least Squares (OLS) regression analysis. Before discussing the results of regression analysis for other variables, this paper elaborates on the findings related to sentiment in the whitepaper.

After reducing the number of independent variables, two observation sets are used in OLS regression analysis on the dependent variable of log return after six months. Table 3.6 shows the result of the OLS regression of both two datasets. The regression result of all 391 tokens generates an adjusted R-squared of 11.6% with five statistically significant variables for the token returns. These five variables are the availability of pre-ICO, ICO duration, token listing duration, market size and the number of restricted areas. The availability of pre-selling tokens had the highest negative coefficient value of -0.5421 compared with other variables. Contrary to expectations, the variables of interest, whitepaper disclosure and the ratio of positive

Table 3.6: The OLS regression results of the relationship between ICO attributes used in this study and token return in the next six months after the token is listed on the exchange. Model 1 took all the 391 tokens, and model 2 only captures 263 tokens that have discrepancies between the token rating given by ICObench and the token return.

	Log Return 6 months	
	(1)	(2)
Independent Variables:		
Whitelist/KYC	-0.1975	-0.304
	(0.201)	(0.186)
Pre-ICO	-0.5421***	-0.4796***
	(0.187)	(0.176)
IEO	-0.3596	-0.3085
	(0.283)	(0.27)
Bonus	-0.3007	-0.1805
	(0.188)	(0.18)
ICO Duration	0.0018*	0.0011
	(0.001)	(0.001)
Token Listing Duration	0.0009**	0.0006
	(0)	(0)
ETH-based	0.034	-0.2445
	(0.262)	(0.258)
Team Size	-0.0061	-0.0042
	(0.009)	(0.008)
CEO	-0.1463	0.0742
	(0.212)	(0.21)
CTO	0.03	-0.0901
	(0.194)	(0.184)
CEO Prev Experience	-0.3647	-0.0753
	(0.659)	(0.609)
Whitepaper Disclosure	0.2904	0.1479
	(0.176)	(0.173)
% Positive Tweets	0.7161	1.2563*
	(0.727)	(0.722)
Market Size	0.0275***	0.0143
	(0.008)	(0.009)
Restricted Area	-0.0405**	-0.0236
	(0.017)	(0.015)
Control Variable:		
Country	-0.015***	-0.0112**
	(0.005)	(0.005)
Token Type	0.0222	0.0079
	(0.022)	(0.022)
Obersvations	391	263
Adjusted R^2	0.116	0.078
Within R^2	0.155	0.138

sentiment in tweets, are not statistically significant predictors of token return. The existence of the whitepaper gave the coefficient value of 0.2904, and a ratio of positive sentiment indicates a strong positive value of 0.7161.

The results from the analysis of 263 ventures with a discrepancy between the rating and token return reveal an adjusted R-squared value of 7.8%, indicating a low contribution of variables to log return. Unlike the overall token dataset, only two variables —pre-ICO, and a ratio of positive sentiment in tweets—are statistically significant to the dependent variable of token return. Pre-ICO had the highest negative effect of -0.4796, while the ratio of positive sentiment in tweets shows a strong positive value of 1.2563, signifying statistical significance to token return. Some variables had opposite coefficient signs compared to the regression result with all ICO ventures, including ETH-based, the existence of a CEO and a CTO; hence, the interpretation of the model coefficient is different. The ETH-based variable has a stronger negative coefficient value of -0.2445, but a weaker value of CEO and CTO existence. However, these three variables are non-statistically significant. Despite the indication of three ICO variables generating opposite signs in two OLS regression models using all ICO ventures and the discrepancy dataset, it is challenging to conclude the set of factors explaining the discrepancy between ICO rating and token value due to their insignificance.

The results presented in Section 3.5.2 confirm the existence of a discrepancy between ICO ratings and token returns. This discrepancy underscores that a favourable ICO rating does not consistently result in a positive long-term financial return, and conversely, a lower rating doesn't always lead to a negative return. Over 67% of the 391 tokens are misclassified based on token ratings from ICObench, highlighting the unreliability of these ratings as a signal for predicting future returns. Additionally, the regression analysis conducted on token returns indicates that ICO ratings have statistical significance, with a negative coefficient of -0.254. The finding supports the initial hypothesis H_1, suggesting that ICO ratings may not be reliable indicators of token performance. This confirmation aligns with previous findings that point out weaknesses in token ratings, such as lack of transparency, unreliable data sources, and the possibility of receiving incentives from ventures to manipulate ratings (Ofir and Sadeh, 2020; Florysiak and Schandlbauer, 2019; Bourveau et al., 2022), despite Lee et al. (2021)'s finding of an association between aggregated ratings and returns.

To address the research questions regarding factors explaining the discrepancy between ICO ratings and token returns, regression analysis is conducted, and results are compared between two groups of observations: all ICO tokens and ICOs with discrepancies. Unfortunately, the statistical results do not provide clear insights into the factors contributing to this mismatch phenomenon. Regression analysis indicates that a few ICO venture characteristics had different directional signs, including the indication of ETH-based tokens and the existence of a CEO and CTO. However, these variables are not statistically significant in predicting token returns.

Possible reasons for these inconclusive results include the dataset size, which may not be sufficient to draw strong evidence. Additionally, ICO ratings rely on information disclosed during the ICO and do not consider post-fundraising information. Moreover, market participants may be focusing on token prices on exchanges rather than the characteristics outlined during the ICO period, and investigating speculative activity in token prices falls outside the scope of this study, but it can be an opportunity for future research.

3.5.4 ICO index result

Table 3.7 presents the outcomes of the regression analysis incorporating the novel ICO index, as detailed in Section 3.4.3, with clustered standard errors based on platform terms. Most independent variables, excluding team attributes, demonstrate statistical significance with respect to six-month token returns. While human capital, such as founders and the team, is widely recognised as a strong contributor to ICO ventures in literature (Campino et al., 2022; Deng et al., 2018; Amsden and Schweizer, 2018; Howell et al., 2019), the analysis fails to confirm hypothesis H_{2b}, as team attributes are not statistically significant for token returns. This suggests that while a strong founding team contributes to the initial success of ICO ventures, its direct positive impact on long-term investment returns may not be statistically evident. Conversely, and in contrast with hypothesis H_{2a}, technology characteristics exhibit a negative and statistically significant effect on ICO returns. This finding suggests that once an ICO is launched and its tokens are traded, information regarding the project's technological characteristics might introduce additional complexity, inadvertently increasing (rather than decreasing) the perceived information asymmetry for investors.

Market characteristics of the token sale, however, yield a positive and statistically significant coefficient, thereby confirming hypothesis H_{2c}. This result aligns with previous findings by Amsden and Schweizer (2018) and Davydiuk et al. (2023), indicating that robust market-related information can narrow the bid-ask spread and contribute to ICO success. Finally, the Twitter sentiment attribute, representing social media, records the strongest positive coefficient value, indicating a significant impact on six-month token returns. This finding confirms hypothesis H_3, underscoring that social media sentiment is a crucial, if not paramount, driver of ICO returns and cannot be neglected in assessment. This result extends previous work, which often focuses on social media activity primarily during the pre-ICO phase (Stanley, 2019).

This study further compares the disclosure index constructed by Bourveau et al. (2022) to evaluate its association with six-month token returns. Previously, Bourveau et al. (2022) create a voluntary disclosure index based on information from whitepapers and other external sources, including source code and social media,

Table 3.7: The OLS regression results of the relationship between the ICO index used in this study and token return in the next six months after the token is listed on the exchange.

	Log Return 6 months				
	(1)	(2)	(3)	(4)	(5)
Independent Variable:					
Team	-0.1474*** (0.042)				-0.0599 (0.04)
Technology		-0.2357*** (0.029)			-0.2167*** (0.027)
Market			0.5425*** (0.06)		0.4969*** (0.059)
Social Media				0.2458 (0.246)	0.6626** (0.289)
Control Variable:					
Country	-0.0104*** (0.002)	-0.0104*** (0.002)	-0.0118*** (0.002)	-0.0092*** (0.002)	-0.0123*** (0.002)
Token Type	0.0394 (0.031)	0.0271 (0.031)	0.0381 (0.028)	0.0428 (0.032)	0.0222 (0.027)
Observation	391	391	391	391	391
Adjusted R^2	0.019	0.047	0.049	0.012	0.075
Within R^2	0.027	0.055	0.056	0.019	0.09

to investigate ICO investment decisions and quantities. For comparative purposes, a modified disclosure index is constructed based on information presented in the whitepaper, following the aspects mentioned in Table 3.1, rather than using sentiment analysis, and then compared with the novel ICO index proposed in this study.

Table 3.8 shows results that indicate the problem description of the ICO project has a strong positive coefficient on six-month token returns, while the description of risks involved exhibits a strong negative coefficient. This suggests the problem statement of the ICO ventures is impactful, and the explicit presence of risks related to investment and product development is inversely associated with returns. Information concerning the team and roadmap presented in the whitepaper also shows a negative coefficient. The remaining whitepaper aspects and the overall length of the whitepaper are not significantly associated with long-term returns. These results contradict Bourveau et al. (2022)'s findings that whitepaper length and technical information are associated with token financing. When these whitepaper indicators are summed as a general disclosure index, it presents a weak positive coefficient that is not statistically significant for the return. In contrast, the proposed ICO index (which is the sum of ICO index categories) shows a negative and statistically significant association with the return.

This study successfully captures sentiment and disclosure signals, constructing a novel ICO index that demonstrates the statistical significance of these signals in influencing the six-month post-ICO return. Interestingly, the results reveal a negative coefficient for team attributes, contradicting assumptions about their critical role in the success of early-stage startups. This trend is similarly reflected in the context of technology attributes. The analysis suggests that the technological capabilities established during token financing may no longer serve as a fundamental driver of price and return once the token becomes available in the market. The proposed ICO index is an attempt to provide investors with a wider and more effective information set to support their financial decision-making. The main point is that even with no further regulatory requirements to the issuers in terms of information provision, it is possible to develop better solutions in the best interest of the individual investors, fundraisers, and the overall financial system level. The analysis enables the decomposition of the information categorised by type. This allows assessment of the contribution from individual components, potentially aiding regulators in minimising the impact of potential new provisions on issuers. Similarly, the findings of this study regarding sentiment as a driver of ICO returns align with the notion that sentiment is a relevant building block in the investor information set and should not be neglected. Prior research suggests that increasing the information available and its reliability improves outcomes in established and standard financial markets, including equity financing. Studies show that increasing the information available and its reliability improves outcomes in established and standard financial markets (and particularly for other equity financing sources). These findings suggest that prioritising reliable information disclosure can be a beneficial approach for the ICO market. The robustness of this novel ICO index in predicting future returns undergo further examination in a machine learning model.

Table 3.8: The OLS regression results of the relationship between token returns six months and two key indices: the Disclosure Index created by Bourveau et al. (2022), and the ICO index developed in this study.

	Log Return 6 months		
	(1)	(2)	(3)
Independent Variable:			
Whitepaper Length	-0.0497 (0.031)		
Problem Description	0.5756*** (0.184)		
Technical	0.3114 (0.279)		
Team	-0.3005* (0.156)		
Roadmap	-0.2656* (0.140)		
Finance	-0.1959 (0.143)		
Business Landscape	0.0760 (0.135)		
Risk	-0.5491** (0.237)		
Disclosure Index		0.0126 (0.022)	
Our ICO Index			-0.1093*** (0.024)
Control Variable:			
Country	-0.0108*** (0.002)	-0.0094*** (0.002)	-0.0102*** (0.002)
Token Type	0.0408 (0.020)	0.0428 (0.031)	0.0340 (0.031)
Observations	391	391	391
Adjusted R^2	0.025	0.012	0.026
Within R^2	0.048	0.019	0.034

3.5.5 Whitepaper sentiment result

Another objective is to investigate the sentiment of the whitepaper, most tokens do not have the whitepaper, limiting the available data to reduce information asymmetry with investors. Additionally, each whitepaper does not necessarily contain all aspects outlined in the IPO prospectus. This incomplete data poses a challenge in conducting OLS regression with other variables. An OLS regression analysis is conducted with ICO returns as the dependent variable and each whitepaper aspect as an independent variable. Only four out of seven categories are found to be statistically significant predictors of token financial returns (as shown in Table 3.9). Although these variables are statistically significant, the negative or low adjusted R-squared values indicate a poorly fitting model. Overall, the sentiment of the whitepaper is not sufficient to confirm the association with the token financial returns in the OLS regression.

Table 3.9: The OLS regression results of the whitepaper in each category and the token return in the next six months after the token is listed on the exchange.

	Log Return 6 months						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Independent Variable							
Problem Description	-0.4141 (0.566)						
Technical		0.338 (0.582)					
Team			0.8732 (0.836)				
Roadmap				-0.9626* (0.509)			
Finance					-0.0104*** (0.002)		
Business Landscape						-0.0104*** (0.002)	
Risk							-0.0104*** (0.002)
Control Variable:							
Country	-0.0086 (0.008)	-0.0074 (0.008)	-0.0024 (0.01)	-0.0019 (0.01)	0.0271 (0.031)	0.0271 (0.031)	0.0271 (0.031)
Token Type	0.0335 (0.031)	0.0299 (0.03)	0.0316 (0.035)	0.0217 (0.034)	-0.2357*** (0.029)	-0.2357*** (0.029)	-0.2357*** (0.029)
Observations	154	158	97	89	102	74	33
Adjusted R^2	-0.003	-0.004	-0.01	-0.018	-0.02	0.006	0.014
Within R^2	0.017	0.015	0.021	0.005	0.01	0.047	0.106

Building upon and extending Bourveau et al. (2022)'s methodology, sentiment analysis is conducted on each whitepaper aspect. The findings align with studies by Florysiak and Schandlbauer (2019); Ante et al. (2018), suggesting that investors may only consider the presence of a whitepaper itself as a positive signal, regardless of the detailed information within or the token return timeframe after fundraising. The association between the existence of a whitepaper and the amount raised in ICO fundraising are confirmed in other literature Ante et al. (2018); Fisch (2019); Bourveau et al. (2022). Therefore, it is not practical to extract sentiment from the whitepaper, given that none of the variables is statistically significant to token return. The findings additionally suggest that the nature of the whitepaper itself may be a contributing factor. Whitepapers may be perceived as technical appendices, potentially limiting their effectiveness within a financial decision-making context. This warrants further investigation.

There are several possible reasons explaining why the sentiment analysis of the whitepaper fails. Firstly, the nature of the whitepaper document may not express emotions as compared to social media text. The result from VADER produced an imbalanced outcome, with many ICOs showing a positive sentiment in each whitepaper aspect. This imbalance lacks negative sentiment to ascertain whether positive or negative sentiment influences future returns. Table 5 also lists the top ten words found in the whitepaper by each category, such as blockchain, data, token, and user are commonly used across various whitepaper aspects and do not convey specific emotions that can be classified into neutral sentiments. VADER sentiment output is analysed, categorised into positive, neutral, and negative sentiments. The average neural sentiment in whitepaper prospectuses is only 2.65%, with the roadmap section comprising 10.2% neural sentiment, while the risk factor lacks any neural sentiment. This underscores the difficulty of capturing sentiment in non-social media text datasets with linguistic characteristics that limit emotional expression. Secondly, the lack of available whitepapers for ICOs poses a challenge; only 161 out of 391 ICOs have whitepapers for analysis, and each whitepaper does not necessarily cover all information categories. Lastly, the application of VADER to extract sentiment from the text can be another limitation. VADER is extensively studied on social media text, but its application to non-social media documents may not be as effective. As a result, these challenges may contribute to why the whitepaper category variable does not provide sufficient information for this study.

3.5.6 Twitter sentiment result

The next notable finding pertains to the sentiment of tweets, specifically the ratio of positive tweets published by token issuers. The study reveals that the sentiment or social attribute of the ICO campaign, constructed as the novel ICO index, is statistically significant with the six-month token return. These regression results indicate the importance of the sentiment analysis of tweets to the future token return, even though, there are concerns about the dataset. The tweets used in this study are extremely unbalanced, with 92.0% of them exhibiting positive sentiments. This imbalance may suggest that the team's objective is to promote products and services while presenting a positive image of the token, potentially avoiding discussions about product risks that may discourage investors. The average compound score of 17,520 tweets is 0.245, indicating a neural sentiment in the three-class classification of tweets, as the value is less than 0.5. Meanwhile, the whitepaper has an average compound score of 0.795, which indicates strong positive sentiment, unlike tweets. The impact of including tweets in predicting token returns is investigated in the following section.

3.6 Classification model

3.6.1 Classification methodology

The last objective of the study involves examining the capability of the machine learning model to predict the six-month post-ICO return. This is framed as a classification problem, where the model predicts whether a token would yield a positive or negative return in the next six months based on the data shared during the ICO fundraising. The dataset is unbalanced, with 118 tokens producing a positive return and 273 tokens experiencing a negative return. To address this classification problem, ICO variables (A) and ICO index (Table 3.3) are utilised as input for the supervised machine learning model to predict the target variable. The decision tree model is chosen as the primary supervised learning approach for its simplicity and capability for exploratory analysis of variables (Myles et al., 2004). Decision trees provide a graphical representation of how the model makes decisions at each node based on input features. This transparency and interpretability stand in contrast to token ratings, where the methodology for computing scores is often not publicly accessible.

A decision tree employs the divide-and-conquer approach, where each tree path contains classification rules assigning the class labels at each node, starting from the root node and reaching the leaf nodes. The scikit-learn library Pedregosa et al. (2011) is used to implement this tree-based model, employing the Classification and Regression Tree (CART) algorithm for training. The CART algorithm iteratively splits the training set into two subsets to identify the pair (k, t_k), where k is a single feature, and t_k producing the purest subsets where all training instances belong to the same class (Géron, 2017). The goal of the CART algorithm is to minimise the cost function (Equation 3.3), utilising impurity measurement on both left and right subsets of the tree. Impurity can be assessed using Gini impurity (Formula 3.4) or entropy (Formula 3.5) as the splitting strategy (Géron, 2017). The algorithm continues splitting the subset recursively until it identifies the best pair, reaches the maximum depth threshold, or no longer finds the best split that reduces impurity or entropy.

$$J(k,t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right} \tag{3.3}$$

where:

$G_{left/right}$ = the impurity of the left/right subset.

$m_{left/right}$ = the number of instances in the left/right subset.

$$G_i = 1 - \sum_{k=1}^{n} (p_{i,k})^2 \tag{3.4}$$

$$H_i = - \sum_{k=1}^{n} (p_{i,k}) log(p_{i,k}) \tag{3.5}$$

where:

$p_{i,k}$ = the ratio of class k instances among the training instances in the class i node.

The decision tree is used for exploratory data and can be utilised in classification and regression problems (Myles et al., 2004). However, this classification model is sensitive to overfitting, which can be prevented by regularisation to limit the degree of freedom. The technique involves splitting training and testing datasets, and pre-determining hyperparameters before training the model. A dataset is divided into training and testing datasets in a 70:30 ratio. Identifying the best combination of hyperparameters for the decision model involves a grid search, evaluating all possible combinations at the expense of high computational time (Chawla et al., 2002). Hyperparameters in decision trees include the choice of impurity criterion, maximum tree depth, the number of samples required to split, and the number of minimum samples on the leaf node (Pedregosa et al., 2011). The prediction model underwent fine-tuning of hyperparameters (Table 3.10) using K-fold cross-validation. This validation splits the training dataset into k subsamples and iteratively trains and validates the hyperparameters search. The F1 score, a mean of precision and recall, considering false positives and false negatives, evaluated the model during the hyperparameter tuning process. To assess overall prediction performance, the F1 score and accuracy, which only considers corrected predictions, are employed.

Table 3.10: The list of hyperparameters of the decision tree model is fine-tuned on the training dataset using grid search to find the best combination of values. The criterion is the function that the CART cost function uses to calculate the impurity of the subset to split. max_depth is the maximum depth of the decision tree. min_sample_leaf is the minimum number of instances on the leaf node, and min_sample_split is the minimum number of samples required to split (Pedregosa et al., 2011).

Model	F1 Score	criterion	max_depth	min_sample_leaf	min_sample_split
Model 1	0.383	entropy	7	3	6
Model 2	0.377	entropy	7	1	9
Model 3	0.307	entropy	7	1	5
Model 4	0.289	entropy	5	3	2

Additionally, the decision tree provides feature importance, quantifying the total decrease in node impurity and indicating the significance of each feature to the target variable.

3.6.2 Prediction result

Figure 3.2 offers valuable insights into the binary prediction of six-month token returns (positive or negative) using a decision tree model. The top panel of the figure illustrates the relative importance of individual ICO variables. ICO fundraising duration stands out as the most significant feature, with a relative importance value of 100. This indicates that the duration of token fundraising plays a critical role in reducing uncertainty when classifying token returns. The ratio of positive sentiment in tweets secures the second position, with a value of 73.4, highlighting the substantial influence of social media perception. Following this, the country of token issuance emerges as the third most important indicator, suggesting that the success of blockchain projects is often tied to their operational jurisdiction.

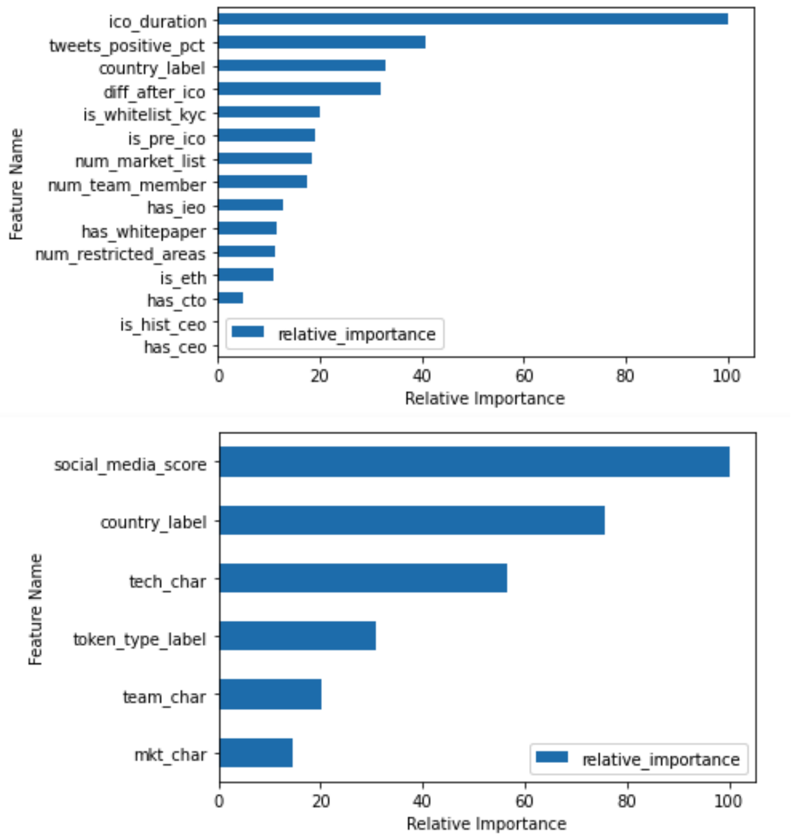

Figure 3.2: The relative importance of decision tree models.

Within the same decision tree model, when incorporating features from the novel ICO index (as detailed in Section 3.4.3), the sentiment score derived from this index achieves the highest relative importance value of 100 among other score variables. This strongly underscores the critical role of aggregated sentiment in influencing the model's predictions, particularly when captured through the structured ICO index. Interestingly, while ICO duration is critical as an individual variable, its prominence diminishes when transformed into a binary feature (below/above median) within the broader technology and campaign categories, becoming less influential compared to the social media attributes.

The classification results for the binary classes of token return are presented in Table 3.11. When examining predictions based on individual features, Model 1, which incorporates tweet sentiment, shows a significant divergence from Model 2, which excludes it. Both models effectively predicted training data, achieving over 85% accuracy. However, Model 1 exhibits difficulties classifying tokens with negative returns, leading to a higher number of false negatives and a reduced overall F1 score compared to Model 2 on the training set. This performance divergence became more apparent on the testing dataset, where the inclusion of the sentiment predictor resulted in a 5.9% higher accuracy rate and 5.3% higher F1 score. This underscores the critical role of social media sentiment in enhancing predictive power, particularly for unseen data.

Table 3.11: The prediction result of token return using the decision tree model. Model 1 uses the same set of ICO variables as the OLS regression model, and Model 2 excludes a ratio of positive tweets.

	Train			Test		
	Confusion Matrix	Accuracy	F1	Confusion Matrix	Accuracy	F1
Model 1	[[184 8] [26 55]]	87.5%	87.1%	[[69 12] [21 16]]	72.0%	70.8%
Model 2	[[175 17] [15 66]]	88.3%	88.3%	[[63 18] [22 15]]	66.1%	65.5%

Furthermore, the inclusion of tweet sentiment also improves the model that uses the ICO index as input, as detailed in Table 3.12. Model 3, which incorporates the tweet sentiment ratio, shows a modest 1.7% improvement in its F1 score on the testing dataset compared to Model 4 (without sentiment), with similar improvements observed in the training datasets. However, models relying on individual variables prove more effective than those using the aggregated ICO index. Specifically, Model 1, leveraging individual features, achieves an F1 score of 70.8% on the testing dataset, outperforming Model 3 (using the ICO index), which achieves an F1 score of 68.5%.

Table 3.12: The prediction result of token return using the decision tree model. Model 3 uses the same set of variables that are grouped into the category of ICO index as the OLS regression model, and Model 4 excludes a ratio of positive tweets.

	Train			Test		
	Confusion Matrix	Accuracy	F1	Confusion Matrix	Accuracy	F1
Model 3	[[188 4] [33 48]]	86.4%	85.4%	[[72 9] [25 12]]	71.2%	68.5%
Model 4	[[182 10] [49 32]]	78.4%	76.0%	[[75 6] [28 9]]	71.2%	66.8%

In summary, the inclusion of tweet sentiments as a predictor further enhances the performance of both models. These promising findings suggest that the prediction model can offer an alternative approach for evaluating ICO tokens, delivering improved transparency and accuracy compared to the ICO ratings. This approach would mitigate the observed discrepancy between ratings and token returns. However, it is important to acknowledge that some misclassification occurs, necessitating further investigation into potential overfitting issues associated with the decision tree's sensitivity to training data.

3.7 Conclusion

Accessing and predicting token returns based on information disclosed during fundraising poses significant challenges. This study explores various signals contributing to token returns, engaging in the ongoing discourse regarding the validity and reliability of token ratings. Notably, ICO ratings misclassified 67% of tokens as positive, prompting an exploration of factors contributing to the discrepancy between ICO ratings and token returns. The regression analysis identifies three venture characteristics —an indication of an ETH-based token, and the presence of a CEO and CTO with contradictory directions, yet their statistical insignificance prevents conclusive explanations for this phenomenon. The limitation can be addressed by utilising a more extensive dataset unrestricted by Twitter presence. Certainly, future research can delve into exploring the relationship between token returns and assessments from alternative token ratings providers. This approach would provide valuable insights into the consistency and reliability of different rating sources in predicting token performance.

In addressing the second objective, focusing on the impact of information disclosure from the whitepaper and Twitter on token returns, the analysis extends beyond overall sentiment, dissecting the whitepaper into sections aligned with the IPO prospectus. Nevertheless, relying on whitepaper sentiment proved impractical, consistent with literature indicating that details, page length, readability, and technical aspects do not significantly influence investment decisions and token returns. Additionally, the study highlighted VADER's limitations in sentiment analysis for non-social media texts, emphasising the need for tailored approaches.

To overcome the transparency issues of token ratings, this study constructs a novel ICO index, demonstrating its significant influence on six-month token returns. The sentiment of tweets, integrated into the ICO index, displays a positive coefficient, underscoring the pivotal role of social media in mitigating information asymmetry and influencing token returns. However, an imbalance in sentiment in tweets suggests a need for further analysis of neutral tweets and exploration of other social media platforms like Facebook and Reddit.

The final contribution involves implementing a machine learning model for predicting future returns, achieving approximately 71% accuracy. Augmenting the model with sentiment from tweets and using the ICO index as an input enhances predictive accuracy. This showcases the potential of machine learning to offer transparent token assessments, surpassing the limitations of token ratings. Future research avenues can explore alternative machine learning models for improved predictions.

In summary, this research underscores the opportunities presented by natural language processing and machine learning in evaluating token fair value. Emphasising the importance of social media signals, the findings encourage entrepreneurs to communicate effectively during blockchain startup fundraising. Investors can leverage artificial intelligence for more accurate token assessments. However, the absence of regulatory oversight for information disclosure necessitates caution, as inaccurate information may impact prediction performance. The study supports the call for regulatory bodies to facilitate transparent voluntary disclosure, fostering promising blockchain projects while mitigating investor biases.

Chapter 4

Startup Valuation with Sustainability: A Novel Approach with Machine Learning and Natural Language Processing

4.1 Introduction

A heightened awareness of environmental protection, social responsibility, and ethical governance (ESG) significantly impacts financial markets, encouraging investors to become early adopters of this emerging trend. This is reflected in the substantial growth of sustainable assets under management (AUM), reaching $35.3 trillion globally by 2020, a 15% increase since 2018 (Global Sustainable Investment Alliance, 2021). Paralleling the well-established research on ESG integration and asset pricing in public equities (Serafeim and Yoon, 2023, 2022; Shanaev and Ghimire, 2022; Gibson Brandon et al., 2021), this trend is now emerging in the private equity market, which focuses on illiquid assets. Sustainability can affect access to finance and the cost of capital for startups and mature private companies (Vismara, 2019; Calic and Mosakowski, 2016; Döll et al., 2022; Hegeman and Sørheim, 2021; Bianchini and Croce, 2022). Consequently, venture capital and private equity firms are integrating sustainability frameworks into their portfolios in two main forms: (1) incorporating sustainability into existing funds or (2) establishing dedicated sustainability-focused funds (Lin, 2022). However, practitioners face significant challenges related to both sustainability investing and the unique characteristics of private capital markets.

Investing in young companies presents a unique risk profile. High failure rates and the inherently illiquid nature of these investments, typically locked in for periods ranging from five to ten years, necessitate a robust valuation process (U.S. Bureau of Labor Statistics, 2022). However, accurate valuation is often hampered by limited data availability. Traditional financial statements and comparable market data points may be ineffective for these companies, potentially leading to valuation errors and impacting the cost of capital. This challenge is further amplified in the context of sustainable investing, where non-financial information such as sustainability reports plays a crucial role in assessing a company's ESG performance. However, resource-constrained early-stage companies may prioritise survival over comprehensive ESG reporting (Lin, 2022). The scarcity of available sustainability data creates a significant impediment for investors seeking to integrate ESG factors into their valuation models.

While third-party sustainability ratings offer a potential solution, their limitations mirror those encountered in the public market. These limitations include the complexity of sustainability terminology and the lack of standardised frameworks for ESG reporting and rating (Boffo and Patalano, 2020; Berg et al., 2022; Gibson Brandon et al., 2021). Furthermore, such sustainability data is often concentrated amongst publicly listed companies, limiting its applicability to the private capital space. To address these shortcomings and effectively assess the value and sustainability of young companies, alternative datasets and technological solutions are gaining traction as methods for evaluating investment opportunities.

The application of artificial intelligence (AI) and its subfield, machine learning, becomes a prominent area of research in financial markets, attracting significant interest from both practitioners and academics. One promising avenue lies in leveraging alternative data sources, like financial news, for textual analysis using Natural Language Processing (NLP) techniques. NLP empowers the model to understand and analyse unstructured text datasets, performing tasks like text similarity analysis and classification. While prior research demonstrates the potential of ML and NLP models for assessing sustainability in public equities (Guo et al., 2020; Ruberg et al., 2021; Gutierrez-Bustamante and Espinosa-Leal, 2022; Mukherjee, 2020; Nugent et al., 2021), the application to startup valuation, particularly for a sustainability context, remains underexplored. This presents a significant research opportunity to leverage sustainability information to enhance the valuation of early-stage startups and inform portfolio investment decisions.

This study addresses a critical research gap: the limited application of machine learning and alternative data sources for early-stage startup valuation, particularly within the field of sustainable investments. A novel machine learning-based valuation model is proposed that conceptually advances traditional methodologies by investigating the impact of sustainability-related textual data on company valuation. This represents a significant contribution, demonstrating a substantial 16.45% improvement in prediction accuracy over conventional approaches.

This research uniquely bridges the domains of sustainable finance and quantitative valuation, providing practitioners in the private capital market with an innovative framework to consider sustainability's crucial, long-term environmental and social impact alongside short-term financial gains. Furthermore, this study offers a conceptual model for integrating complex, unstructured data, aligning with the growing imperative for enhanced sustainability disclosure advocated by policymakers. Through this integration, it not only contributes to academic literature with enhanced valuation techniques but also provides practical impetus for more comprehensive and sustainable investment decision-making in venture capital and private equity.

The remainder of this chapter is organised as follows. Section 2 outlines the background of startup valuation in venture capital investment, sustainable investment in public, and machine learning applications in finance. Section 3 details the employed methodologies, including machine learning models and natural language processing. Section 4 introduces the dataset used in this study. Section 5 presents the study's findings, while Section 6 discusses these findings, their implications, and any limitations of the research. Finally, Section 7 concludes the paper and outlines potential areas for future work.

4.2 Literature review

4.2.1 Startup valuation methodology

Successfully navigating the early-stage life cycle is critical for young companies. This phase is characterised by high upfront costs associated with product development and market entry before generating revenue. Equity financing, often secured from angel investors or venture capitalists (VCs), provides crucial funding in exchange for company ownership. Determining the appropriate valuation at this stage is essential, as it dictates the equity price, and funding potential becomes paramount for investors seeking high returns despite the inherent risk. Startups face high failure rates (i.e., 20% fail within the first year (U.S. Bureau of Labor Statistics, 2022)) and illiquid investments, further amplifying risk. Additional complexities in startup valuation arise from market uncertainty, unique business models, multi-stage financing rounds, limited financial data, and information asymmetry (Sander and Kõomägi, 2007; Montani et al., 2020; Damodaran, 2009). The lack of legal obligation for private companies to disclose financial and non-financial information creates a particularly high level of information asymmetry (Montani et al., 2020). Consequently, VCs heavily rely on their own experience when making investment decisions, potentially leading to discrepancies in agreements with entrepreneurs (Glücksman, 2020). These factors collectively present significant challenges for company valuation within the private capital market.

Traditional valuation methods, such as discounted cash flow (DCF), rely on predictable cash flows and established discount rates to estimate a company's present value (Williams, 1938). However, the inherent uncertainty associated with early-stage startups poses significant challenges for applying DCF. The lack of historical financial records and the absence of positive earnings in the early stages further complicate the process (Damodaran, 2009; Sander and Kõomägi, 2007). In response to these limitations, researchers propose incorporating non-financial information and growth potential into valuation frameworks (Myers, 1977; Black, 2003; Shevlin, 1996; Köhn, 2018). For instance, Black (2003) suggests using the incremental cash flows of unrecognised net assets to quantify growth opportunities in young firms.

Damodaran (2009) outlines both top-down and bottom-up approaches for estimating cash flows and discount rates considering the market risk and the correlation of the VC portfolio. Additionally, Damodaran (2009) proposes a "key person discount" to account for the potential impact of losing key personnel on a company's success, earnings, and cash flow generation. This concept is supported by other studies that highlight the crucial role founders and teams play in successful fundraising and valuation increases (Macmillan et al., 1985; Miloud et al., 2012; Hsu, 2007).

Beyond traditional valuation methods, VCs sometimes utilise alternative approaches like the Berkus method (Payne, 2011). This model assigns pre-defined financial values (up to $0.5 million) to a set of qualitative factors that represent common risk factors for startups. These factors may include the experience of the management team in the relevant industry, the soundness of the business idea, the existence of a working prototype, and the progress of product rollout or sales. However, the Berkus method's simplicity introduces a significant limitation: its reliance on a fixed set of five factors may overlook industry-specific elements crucial to a startup's success, such as regulatory or technological disruptions, potentially leading to inaccurate valuations (Payne, 2011). These limitations highlight the need for valuation frameworks that can incorporate both qualitative and quantitative data, while also being adaptable to the unique risk profiles of startups across different industries.

The choice of valuation approach employed by practitioners is influenced by additional factors. Sander and Kõomägi (2007) highlight the role of geographical location, demonstrating that private equity and venture capital firms may favour different methods based on their region. For example, Estonian investors tend to utilise cash flow-based valuations (e.g., DCF) compared to Western firms that rely more heavily on multiples and comparable metrics. This finding aligns with Reverte et al. (2016), who suggest that VCs in countries with English-based legal systems, like the UK, are more likely to use comparable metrics due to the prevalence of well-established and reliable companies for comparison. However, the unique and non-replicable nature of early-stage business models, coupled with limited market data for such companies, presents a significant challenge in identifying truly comparable companies within the same sector (Montani et al., 2020). This difficulty in finding suitable comparables further underscores the need for alternative valuation tools.

Recent research explore the application of machine learning (ML) models for predicting pre- and post-money valuations of companies (Miloud et al., 2012; Ang et al., 2022; Zhang et al., 2023; Garkavenko et al., 2021). Ang et al. (2022) employ ElasticNet and XGBoost to predict post-money and pre-money valuations, respectively. Notably, their findings suggest a non-linear relationship between company determinants and predicted value, with the amount raised exerting the highest influence on both valuations. Miloud et al. (2012) use linear regression to explore the relationship between pre-money valuation and factors like product differentiation, industry growth, management team experience, and network size. Zhang et al. (2023) propose a custom model called Adam-ENN utilising differential evolution and an adaptive learning rate optimisation algorithm. This model outperforms benchmarks (Least Squares, Ridge Regression, Deep Neural Network, Random Forest) and identifies the number of VC investors as the most significant factor, likely due to their social capital and network contributions to the portfolio company. Garkavenko et al. (2021) develop a Domain Adaptation framework that surpasses EPoSV, CatBoost, and MLP models in performance. Notably, their model identifies financing round information (total amount raised and series type) as the most relevant factors. While Montani et al. (2020) acknowledges the absence of a perfect valuation model for early-stage companies, there remains room for improvement. This study proposes a novel machine learning-based model that integrates the concept of sustainability, further explained in subsequent sections.

4.2.2 Sustainable investment and assessment

Sustainable investing gains significant traction in recent years, driven by several key factors. An OECD report (Boffo and Patalano, 2020) and accompanying survey by BNP (2019) identify the desire for improved long-term returns, enhanced firm reputation, and alignment with social and moral considerations as key drivers of ESG investing. The belief that sustainable portfolios outperform traditional ones further fuels this trend. Institute for Sustainable Investing (2021)'s report highlights that US sustainable equity and bond funds outperform their peers by a median of 3.9% and 2.3%, respectively. Early adoption of sustainability challenges can lead to significant value creation for businesses (Fatemi and Fooladi, 2013). This value creation manifests through enhanced brand reputation, increased customer loyalty, and improved shareholder returns. Conversely, neglecting sustainability can lead to value deconstruction. This dynamic incentivises entrepreneurs to adapt their strategies and operations to achieve sustainable growth and meet evolving stakeholder demands. Beyond investor preferences, policymakers and regulators are actively shaping the sustainability landscape through two key initiatives: (1) promoting economically, environmentally, and socially responsible development (Cumming et al., 2022), and (2) supporting sustainability reporting and responsible investment practices (Global Sustainable Investment Alliance, 2020). Scholars are actively researching this emerging field, analysing the significance of sustainable investing and how practitioners develop and utilise tools to assess sustainability metrics aligned with ESG principles.

Research in sustainable investing primarily concentrates on the relationship between ESG signals and investment returns. One prominent source of these signals is aggregated ESG ratings and indices provided by firms like Sustainalytics and MSCI (e.g., MSCI ESG Ratings, MSCI World SRI Index). Institute investors and fund managers widely utilise these indices to assess portfolio risk exposure and identify potential investment opportunities (Deutsche Bank, 2021). Compared to other ESG-related information sources, such as company sustainability reports, corporate social responsibility (CSR) reports, news articles, filings, and regulatory reports, aggregated

ESG ratings remain the most popular choice. A substantial body of academic research investigates the connection between ESG ratings and various financial aspects of investments. These studies explore the relationship with factors such as pricing, returns, and the cost of debt financing (Serafeim and Yoon, 2023; Gibson Brandon et al., 2021). However, discrepancies between ESG ratings from various providers emerge, raising concerns about the consistency and reliability of these data sources used in investment decisions. This, in turn, heightens scrutiny regarding the potential for “greenwashing", where companies may overstate their ESG commitment.

While ESG ratings offer valuable insights into the sustainability of investment targets, inherent limitations can lead to oversimplification and potentially hinder the achievement of desired financial and social returns. Boffo and Patalano (2020) highlight several key challenges associated with ESG ratings. These include inconsistency across providers, such as MSCI and Sustainalytics, who may assign different weightings to ESG pillars, indicators, and materiality considerations. This inconsistency can lead to significantly different ratings, particularly when comparing companies across industries. Furthermore, the lack of transparency in rating methodologies and data sources used by agencies limits interpretability for users. Finally, ESG ratings often rely on potentially biased or non-standardised ESG disclosures from companies, potentially introducing further inaccuracies. These limitations exacerbate rating divergence across industries (Berg et al., 2019; Rajna, 2021), hindering effective risk and opportunity assessment and potentially causing asset mispricing in sustainability-focused portfolios (Berg et al., 2022; Boffo and Patalano, 2020; Gibson Brandon et al., 2021; Serafeim and Yoon, 2023).

Highlighting a further complication Gibson Brandon et al. (2021) find a particularly strong disagreement between ESG ratings and financial returns, especially within the environmental pillar. Similarly, Serafeim and Yoon (2023) identify that high levels of ESG rating disagreement weaken the ability to predict market reactions and the relationship between news sentiment and financial returns. Finally, the potential for greenwashing, where companies misrepresent their sustainability commitment, requires ongoing attention (Yu et al., 2020). The lack of standardised disclosure practices and robust audit mechanisms further complicates the integration of ESG and sustainability-related factors into investment decisions (Ho, 2015). As a result, practitioners and academics are actively seeking alternative solutions to these challenges.

The limitations of traditional ESG ratings underscore the need for alternative data sources to assess a company's sustainability performance and its impact on financial assets. News articles and sustainability reports emerge as promising alternatives, with studies demonstrating their influence on future asset returns and company risk (Guo et al., 2020; Schmidt, 2019). To unlock the potential of these unstructured text data sources, researchers and practitioners are increasingly turning to artificial intelligence (AI), specifically natural language processing (NLP) techniques.

While significant research explores the application of NLP in the public financial market (Krappel et al., 2021; Guo et al., 2020; Ruberg et al., 2021; Gutierrez-Bustamante and Espinosa-Leal, 2022), a critical gap remains in the private capital space. Existing studies primarily focus on tasks like predicting public company sustainability ratings, and stock volatility, or classifying reports against frameworks like the Global Reporting Initiative (GRI). Recent advancements in NLP models offer promising solutions. For example, domain-specific models called ESG-BERT (Mukherjee, 2020) are pre-trained specifically to categorise text into ESG categories. Additionally, NLP models can be pre-trained on financial corpora and fine-tuned for specific tasks, such as multi-label classification of text related to ESG controversy and UN sustainable Development Goals (SDGs) goals (Nugent et al., 2021). These advancements, coupled with data augmentation techniques to address limitations in labelled data for private markets, hold significant potential for enhancing the accuracy and comprehensiveness of sustainability assessments in private capital.

However, the effectiveness of sustainability reporting remains hampered by a lack of standardised guidelines and robust verification mechanisms. While voluntary initiatives like the UN Sustainable Development Goals (United Nations, 2015) and the Sustainability Accounting Standards Board (SASB) standards (IFRS Foundation, 2022b) aim to address this issue, further regulatory action is necessary. The EU's

Sustainable Finance Disclosure Regulation (SFDR) mandates market participant disclosure of ESG integration processes and data sources, promoting standardisation and mitigating greenwashing risks (European Comission, 2022). While numerous studies explore sustainability investing in public markets, the applicability of these strategies to private capital markets remains an open question.

4.3 Sustainable investment in private capital: Hypothesis development

While long-term investment horizons in venture capital and private equity offer the potential for sustainability-driven value creation. However, a key challenge lies in the lack of mandatory sustainability disclosure by private companies, forcing investors to rely on information gleaned directly from founders and management teams (Lin, 2022). The evaluation of sustainable VC funds can be conducted through two main approaches: (1) adding sustainability criteria to existing funds and (2) launching dedicated sustainability-focused funds (Lin, 2022). Institutional investors can choose either approach, factoring sustainability considerations into their investment selection and execution processes (Lin, 2022; Wiek et al., 2023).

For a successful sustainable investment program, firms must prioritise two key practices: measuring the impact of sustainability initiatives and developing a firm-wide sustainability strategy (Wiek et al., 2023). Despite the emergence of ESG tools like the Preqin database (Preqin, n.d.), which offer access to ESG funds and risk assessment capabilities, key challenges in sustainable investment remain. The lack of standardised data for measuring impact and benchmarking sustainability goals is a concern for investors, particularly for young companies. A key concern persists regarding the potential trade-off between prioritising sustainability initiatives and achieving strong financial returns in venture capital (VC) investments. This concern is further amplified by the inherent uncertainty associated with early-stage investing, where investors, including limited partners (LPs), may encounter unknown risks and market complexities (Lin, 2022).

While challenges exist, the growing importance of sustainability considerations cannot be ignored in the private capital market. Venture capitalists, whose decision-making is extensively studied (Corea et al., 2021a), face an opportunity to be proactive in meeting this evolving demand (Cumming et al., 2022). This necessitates further research on how investors can select portfolios that balance economic, environmental, and social objectives, and how such investments can contribute to sustainable development goals. Furthermore, the growing importance of sustainability is evident in emerging areas of entrepreneurial finance, such as crowdfunding and corporate venture capital (CVC).

Existing literature presents mixed findings regarding the relationship between a firm's sustainability orientation and crowdfunding success. Vismara (2019) investigates the influence of sustainability on equity-based crowdfunding, concluding that it had no significant positive impact on campaign success rates. However, their analysis observes that such firms attract a different investor base, one motivated by factors beyond just financial returns. Conversely, Calic and Mosakowski (2016) report a positive influence of social and environmental sustainability on the funding success of reward-based crowdfunding projects. In addition, Mansouri and Momtaz (2022) examine Initial Coin Offerings (ICOs), characterised as a form of blockchain-based crowdfunding, and find that sustainability orientation (as determined by whitepaper word count) positively correlates with funding acquisition, while conversely, it negatively affects subsequent one-year token returns. These contrasting findings highlight the potential for crowdfunding and investor motivations to moderate the effect of sustainability orientation on campaign outcomes.

Large corporations are increasingly leveraging their CVC arms to invest in sustainable businesses, particularly those in the clean technology (cleantech) sector. This strategy offers a dual benefit: building a competitive advantage through innovation and promoting CSR initiatives, ultimately aligning with long-term strategic and financial objectives (Döll et al., 2022; Hegeman and Sørheim, 2021). While sustainability factors are gaining traction, their impact may vary across industries. Bianchini and Croce (2022) highlight the case of cleantech, where VC firms may be hesitant to invest due to lower perceived returns and space-intensive products compared to sectors like information and communication technology (ICT) and biotechnology.

Prior research highlights the potential impact of sustainability on entrepreneurship financing. However, a critical gap remains in understanding the long-term relationship between sustainability and venture success, particularly in terms of valuation. Filling this gap requires a deeper examination of how a venture's commitment to sustainability translates into startup valuation over time. Building on the potential benefits of sustainability for value creation and long-term market positioning, the study proposes the following hypothesis:

> H_1: There is a positive relationship between the level of sustainability exposure, as measured by media coverage of ESG initiatives, and its valuation. The study assumes that ventures with public exposure to environmental, social, and governance pillars are more attractive to investors and can stimulate market growth, potentially leading to higher valuations.
>
> H_2: There is a non-linear relationship between variables predicting venture valuations, and deep learning models are capable of capturing these complex interactions, potentially leading to a more comprehensive understanding of the factors influencing venture valuation. A deep learning model will outperform a traditional supervised learning model in predicting venture valuation.
>
> H_3: A text embedding model pre-trained on ESG and financial corpus will lead to a more accurate venture's valuation compared to a traditional text embedding model. By incorporating domain-specific information on sustainability, the pre-trained model is expected to capture a more nuanced understanding of the venture's potential and improve prediction accuracy.

News articles offer a promising alternative data source for evaluating a venture's sustainability performance, complementing traditional methods employed by VCs such as founder experience and market size (Arroyo et al., 2019; Corea et al., 2021a). While recent research utilises network analysis on news articles for M&A prediction (Venuti, 2021), a crucial aspect often remains underexplored: sustainability. This study proposes an innovative approach to integrate AI, specifically NLP techniques, to analyse news articles and assess a startup's commitment to sustainability. This proposed approach to sustainability assessment has the potential to revolutionise startup valuation in the private capital market, enabling VCs to make more informed, future-proof investment decisions that consider not only short-term gains but also long-term environmental and social impact.

4.4 Methodology

This study proposes a novel approach to traditional startup valuation by integrating news articles and sustainability context through supervised learning and natural language processing (NLP). These alternative data sources hold promise for capturing valuable insights beyond traditional financial metrics. This section outlines the details of both supervised learning and the NLP techniques used to process and integrate unstructured text data from news articles into the valuation model. The diagram of the novel startup valuation model is illustrated in Figure 4.1, with further details on the data collection provided in Section 4.5.

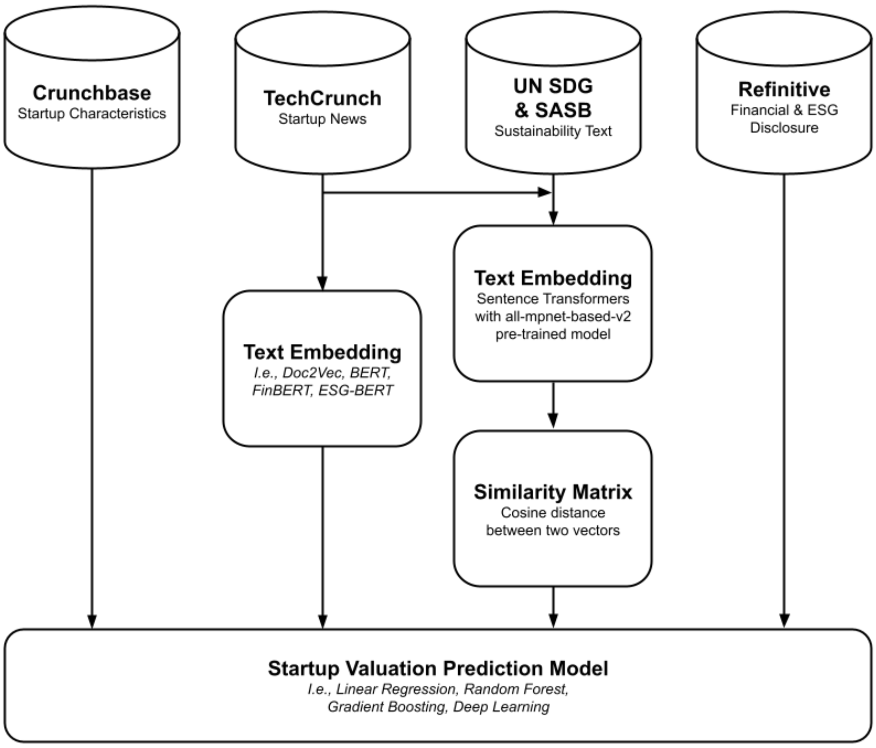

Figure 4.1: A diagram shows the sources of data collection and the methodology used in the novel startup valuation model.

The objective is to predict the pre-money valuation of startups at each funding round, incorporating news articles and their potential influence on valuation. The prediction builds on the assumption that news articles published prior to or in the same year as the deal announcement have the most significant influence on a startup's valuation at that particular funding stage. These news articles likely capture the most recent developments and investor sentiment surrounding the venture, potentially impacting the pre-money valuation assigned during the funding round. The details of the constructed dataset used to test this hypothesis are further explained in the next section.

The study employs a supervised learning approach, where algorithms are trained on labelled data with known target variables. To preserve temporal order and mimic real-world prediction scenarios, a chronological splitting strategy with a simulation window concept is adopted to split the data into training and testing datasets (Arroyo et al., 2019; Corea et al., 2021a). Similar to time series analysis, historical data from funding rounds between 2010 and 2018 is used for training the model, while more recent data (2019-2020) is reserved for validation and testing to assess model performance as shown in Table 4.1. This approach aligns with the reasoning presented in Arroyo et al. (2019) by keeping the simulation window close to the current period. This strategy mitigates survivorship bias[1] and allows the model to learn how startups respond to contemporary sustainability trends, which may differ from those reflected in the past.

Table 4.1: This table details the chronological split of the Crunchbase (CB) and TechCrunch (TC) datasets used for training and testing the valuation prediction model.

Type	Year	CB Size	TC Size
Training	2010-2018	648	5815
Testing	2019-2020	464	2972
	Total	1112	8787

4.4.1 Machine learning baseline models

Given the study's objective of predicting startup valuation as a regression problem, several supervised learning models can be considered. To explore the potential benefits of complex model architectures, the experiment design is divided into two categories: baseline models and deep learning models. The baseline models aim to capture linear relationships within the data. Three common regression algorithms are implemented: Linear Regression (LR), Gradient Boosting (GB), and Random Forest

[1] Companies that fail early before Crunchbase was established may not be available in the database: this can lead to overrepresentation of successful and still-operating companies (Arroyo et al., 2019).

(RF) using the scikit-learn library (Pedregosa et al., 2011). These models provide a benchmark for comparison with the more complex deep learning models.

To ensure optimal performance of prediction models, the hyperparameter tuning through RandomizedSearch and cross-validation is employed. Cross-validation strengthens this process by splitting the data into training and validation sets across multiple iterations. Each parameter combination is assessed on the training data, with its performance measured using Mean Squared Error (MSE) on the validation set (Section 4.4.4). Minimising MSE on the validation set guides the search towards hyperparameter combinations likely to generalise well on unseen data, ultimately enhancing the prediction model's accuracy.

Linear Regression (LR)

Linear regression serves as a foundational model for valuation prediction. It establishes a linear relationship between the independent variables (startup attributes) and the dependent variable (pre-money valuation). While its simplicity offers ease of interpretation, it has limitations in capturing complex, non-linear patterns that may exist within the data. Additionally, the model relies on several key assumptions: absence of multicollinearity (high correlation among independent variables), exogenous property (independence from the error term), and normally distributed error terms with constant variance across all levels of the independent variables. Violations of these assumptions can compromise the model's reliability.

The formula for multiple linear regression, which involves more than one independent variable, is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{4.1}$$

where:

y = A log of pre-money valuation as dependent variable

x$_1, ..., x_n$ = Independent variables

b$_0, b_1, ..., b_n$ = Coefficients represent relationship between dependent and independent

ϵ = Residuals $\sim N(0, \sigma^2)$

Gradient boosting (GB)

Gradient boosting, introduced by Friedman (2001), offers another approach to regression problems. This ensemble method leverages the concept of weak learners, often decision trees, to iteratively build a more robust model. The key idea is to sequentially train these learners, with each one attempting to correct the errors (residuals) of the previous learner in the ensemble.

Each iteration m $(1 \leq m \leq M)$ trains a new model $F_m(x)$ to predict the target variable $(\hat{y})$ as shown in Formula 4.2. This process minimises a loss function, often the mean squared error (MSE) (Section 4.4.4), by fitting the negative gradient of the loss function. This iterative approach strengthens gradient boosting as it progressively refines predictions by focusing on the errors from prior learners. Table 4.2 presents the hyperparameters used in gradient boosting.

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{4.2}$$

where:

F$_m(x)$ = The predicted value at iteration m

F$_{m-1}(x)$ = The predicted value at iteration $m-1$

h$_m(x)$ = The predicted value of weak learner at iteration

ν = A learning rate

Through an iterative process involving M steps, the model progressively refines the prediction. The final prediction is then equivalent to the sum of these individual contributions, as shown in the formula below:

$$F(x) = F_m(x) = F_0(x) + \nu \cdot \sum{}^{M}{}_{m=1} h_m(x) \tag{4.3}$$

Table 4.2: Hyperparameters for Gradient Boosting and Random Forest Models

Model	Parameter	Description	Value
Gradient Boosting	n_estimators	The number of boosting stages or iterations to perform.	[100, 200, 300]
	learning_rate	The learning rate or step size of each boosting iteration.	[0.1, 0.01, 0.001]
	max_depth	The maximum depth of each decision tree in the ensemble.	[3, 5, 7]
	subsample	The fraction of samples to be used for training each individual tree.	[0.8, 1.0]
	min_samples_split	The minimum number of samples required to split an internal node.	[2, 4, 6]
Random Forest	n_estimators	The number of trees.	[100, 200, 300]
	max_depth	The maximum depth of each tree.	[None, 5, 10]
	min_samples_split	The minimum number of samples required to split an internal node.	[2, 5, 10]
	min_samples_leaf	The minimum number of samples required to be at a leaf node.	[1, 2, 4]
	max_features	The number of features to consider when looking for the best split at each tree node.	[’auto’, ’sqrt’]
	bootstrap	Whether bootstrap samples are used when building trees. If True, each tree is built on a random subset of the training data with replacement.	[True, False]

Random forest (RF)

Random forests (Breiman, 2001) represent another ensemble learning technique for valuation prediction. This approach aggregates the predictions from multiple decision trees, leading to a more robust and generalisable model. Each decision tree within the forest is independently constructed, introducing an element of diversity that strengthens the overall model's performance.

During tree construction, the best split at each node is determined by a splitting criterion, often the mean squared error (MSE), which minimises the $L2$ loss (i.e., Least Square Errors). The final prediction for a given startup is determined by aggregating the individual predictions from all trees in the forest. This aggregation can be achieved by averaging (mean) or taking the most frequent prediction (median) across the ensemble (Formula 4.4).

This process allows the model to capture non-linear relationships between independent and dependent variables, potentially offering an advantage over simpler models like linear regression. The hyperparameter configuration used for the random forest model is presented in Table 4.2.

$$y = \frac{1}{N} \sum_{i=1}^{N} f_i(x') \tag{4.4}$$

where:

y = A log value of pre-money valuation as dependent variable

N = The number of decision trees

f$_i(x')$ = The predicted value of each decision tree i with unseen samples x'

Deep learning (DL)

Deep learning is a subset of machine learning that utilises neural networks with multiple layers inspired by the human brain's structure and function. Deep learning is extensively explored in literature for its ability to uncover intricate patterns within large datasets. They consist of interconnected processing units called artificial neurons, which receive signals from other neurons and apply activation functions to produce an output. The network learns complex patterns by iteratively adjusting the weights of connections between neurons. This process is driven by the goal of minimising the cost function, which measures the difference between the predicted and actual values.

The model employed in this study utilises a feed-forward architecture. In this architecture, the input data (i.e., Crunchbase variable and sentence vectors) is passed through each layer of the network. Within each layer, weights are applied to the data, followed by an activation function to introduce non-linearity. A common activation function used in this study is the Rectified Linear Unit (ReLU), which is defined as $f(x) = max(0,x)$, where x represents the input value. Outputs greater than zero remain unchanged, while negative outputs are set to zero.

A cost function, such as the mean squared error (MSE), is used to measure the difference between the predicted and actual values. The model aims to minimise this cost function by iteratively adjusting the weights of the connections between neurons through a process called backpropagation. Backpropagation calculates the gradient of the cost function with respect to each weight, allowing the model to update the weights in a direction that reduces the overall error. This process continues for a set number of epochs (training iterations) or until a satisfactory level of performance is achieved.

The PyTorch library (Paszke et al., 2019), a deep learning framework based on the Torch language, is used to create a custom multi-layer perceptron (MLP) model. The model consists of three linear layers and two activation functions. The Adam optimiser (Appendix C) is employed to optimise the model's learning process. Cross-validation is utilised to identify the optimal hyperparameter combination, including the learning rate (γ) from a range of $[1e-3, 1e-4, 1e-5]$ and the number of epochs from a range of $[5, 10, 50, 100]$. The batch size is fixed at 32. The trained model is then evaluated using a separate testing dataset.

4.4.2 Text embedding models

The study leverages Natural Language Processing (NLP) techniques to process unstructured text data from news articles for integration with machine learning models. To transform the preprocessed text into numerical representations suitable for machine learning algorithms, the study employs two text embedding models. The first model, Doc2Vec (Le and Mikolov, 2014), excels at capturing grammatical structure and contextual information within news articles. The second model utilises pre-trained Large Language Models (LLMs) specifically trained on a corpus of financial and sustainability-related text. These pre-trained LLMs offer the advantage of incorporating domain-specific knowledge into the text representation, potentially leading to more accurate valuation predictions.

Doc2Vec models

The Doc2Vec model (Le and Mikolov, 2014) is a document-level or sentence-level model that transforms textual news articles into numerical representations. Unlike Bag-of-Words models that focus solely on word frequency, Doc2Vec incorporates context to capture the semantic relationships between words.

This study utilises the Distributed Memory (PV-DM) variant of Doc2Vec. PV-DM excels at predicting the next word within a defined context window based on both sentence vectors and content words. This approach enables the model to consider the surrounding context and leverage sentence vectors as a form of memory to retain topic-related information. In contrast, the Distributed Bag-of-Words (PV-DBOW) variant discards context words, relying solely on word representations independent of context and order. This limitation makes PV-DBOW less suitable for capturing the nuances of financial news articles, where contextual understanding is crucial.

The Gensim library (Rehurek and Sojka, 2011) is employed to train the Doc2Vec model on the list of words within each document or sentence, generating corresponding vector representations. This study experiments with different vector output sizes $(50, 100, 150)$ as hyperparameters to assess their impact on valuation predictions within regression models.

Unlike structured features readily available for startups, news text necessitates preprocessing before incorporation into the valuation prediction framework. A widely adopted industry and academic approach is followed. The initial step involves text cleaning, which removes noise such as stop words, punctuation that does not provide semantic value. Subsequently, the text is converted to lowercase and tokenised into individual words. Lemmatisation is then applied to map words to their dictionary-based forms, enabling consistency and reducing vocabulary size. This preprocessing ensures the model focuses on the most relevant content within the news articles.

Large Language Models

Large Language Models (LLMs) are a class of neural network models trained on massive amounts of text data. Through extensive fine-tuning, LLMs achieve remarkable performance on various tasks, including document classification, sentiment analysis, and question answering. This study leverages an LLM as a feature extractor to generate contextualised sentence representations for use in valuation prediction. These representations encode the meaning of a sentence while considering the surrounding context in bidirectional nature (i.e., both before and after), providing valuable features for machine learning models (Devlin et al., 2019). Notably, various pre-trained LLM options exist, including models trained on general corpora like Bidirectional Encoder Representations from Transformers (BERT). The pre-training processes of BERT are self-supervising learning that involves predicting missing words that are masked with the token and next sentence prediction, whether they followed the previous sentence together (Devlin et al., 2019). This study employs FinBERT (Araci, 2019) and ESG-BERT (Mukherjee, 2020), LLMs pre-trained in financial and sustainability corpora, respectively. These domain-specific models potentially offer advantages by incorporating relevant knowledge into the text representation.

- **FinBERT**: Building upon the BERT architecture, FinBERT is pre-trained on a corpus of financial news articles from Reuters (Araci, 2019). This pre-training process potentially imbues the model with domain-specific knowledge relevant to financial tasks. While originally fine-tuned for sentiment analysis, this study explores its applicability in generating sentence embeddings through token averaging using FinBERT embedding model (Kumar, 2021).

- **ESG-BERT**: Developed by Mukherjee (2020), ESG-BERT is designed for text classification related to 26 sustainability topics. Notably, 25 of these topics align with the general issue categories published by the Sustainability Accounting Standards Board (SASB) (see Appendix G for details). This alignment suggests the model's ability to identify relevant sustainability risks and opportunities within text data. However, the specific corpus used for pre-training ESG-BERT remains undisclosed, raising questions about the data's relevance to the financial domain (Ruberg et al., 2021).

Unlike Doc2Vec preprocessing, stopword removal and lemmatisation are omitted to preserve context for these models. The text is converted to lowercase before being embedded by each embedding model into a 768-dimensional vector. While BERT model[2] and ESG-BERT are originally designed for classification tasks, this study adapts them for regression. To achieve this, sentence embeddings are generated by averaging the pooled token embeddings from each model, resulting in a 768-dimensional representation suitable for feeding into the regression models.

4.4.3 Sustainability text similarity

To quantify the relationship between startup news and sustainability principles, this study employs sentence transformers to convert them into numerical representations. Sentence transformers excel at capturing semantic similarity relationships between sentences (Reimers and Gurevych, 2019). This approach draws inspiration from the work of Gutierrez-Bustamante and Espinosa-Leal (2022), who measure the textual similarity between corporate social responsibility (CSR) reports of Nordic companies and the Global Reporting Initiative (GRI) framework. The study leverages the *all-mpnet-based-v2* to encode sentences and short paragraphs. This model achieves the best performance in sentence embedding across 14 datasets, demonstrating significant capability in semantic search (Reimers and Gurevych, 2019).

Following sentence transformation, the cosine distance (Formula 4.5) is calculated between the two sentence embedding vectors. Cosine similarity, a well-

[2] This study uses the pre-trained bert-base-uncased model, originally trained on a corpus of Wikipedia and books (Devlin et al., 2019).

established method in text analytics, is frequently used to measure sentence and document similarity (Reimers and Gurevych, 2019). Appendix F provides examples of news titles exhibiting high similarity to sustainability-related text, including UN SDGs and SASB standards.

$$similarlity = cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum^{n}{}_{i=1} A_i B_i}{\sqrt{\sum^{n}{}_{i=1} A_i^2}\sqrt{\sum^{n}{}_{i=1} B_i^2}} \tag{4.5}$$

4.4.4 Performance metrics

Given the regression nature of the startup valuation prediction task, this study uses a set of four evaluation metrics to assess model performance.

- **Mean squared error (MSE)**: This metric measures the average squared difference between predicted and actual valuation values. MSE is calculated as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y - \bar{y})^2 \tag{4.6}$$

- **R-Squared (R^2)**: R^2 measures the proportion of variance in the actual valuations which reflects the goodness of fit i.e., how well the data fit the regressor. It is calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} \tag{4.7}$$

where SS_{res} is the sum of squared residuals (the squared differences between the observed values and the predicted values) and SS_{total} is the total sum of squared, representing the variance in the observed values. R^2 values closer to 1 indicate a stronger model fit.

- **Root Mean Square Error (RMSE)**: The square root of MSE provides the average magnitude of the errors between predicted and actual valuations in the same units as the original data.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y-\bar{y})^2} \tag{4.8}$$

- **Mean Absolute Error (MAE)**: This metric calculates the average of the absolute differences between predicted and actual valuations. While it is less sensitive to outliers compared to MSE, it does not take the magnitude of the errors into account.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i-\bar{y}_i| \tag{4.9}$$

Lower values for all three metrics (MSE, RMSE, and MAE) indicate a better model fit, minimising the discrepancies between predicted and actual valuation values.

4.5 Data description

This section details the four primary data sources employed in startup valuation prediction. Data sources encompass early-stage funding from the Crunchbase databases capturing the startups' characteristics and history of funding. News articles from the TechCrunch platform provide insights into current developments and sentiment surrounding each startup, potentially influencing investor decisions. Furthermore, data on the Sustainable Development Goals (SDGs) and industry-specific sustainability criteria from the Sustainability Accounting Standards Board (SASB) are incorporated to assess the focus on sustainability within each startup. Finally, the ESG scores obtained from Refinitiv provide a quantitative measure of a startup's commitment to sustainable practices, complementing the other data sources.

Crunchbase

This study leverages data from Crunchbase, a comprehensive database widely used in academic and industry research (Ang et al., 2022; Garkavenko et al., 2021; Krishna et al., 2016). Crunchbase offers detailed information on startups and private companies on a global scale, including portfolio companies, funding rounds, investors, and individuals associated with the startup. A detailed list of variables extracted from Crunchbase and their preprocessing steps for transformation into numerical representations is provided in Appendix D.

TechCrunch

This study complements the structured data with unstructured text data from TechCrunch, a leading online publication that focuses on startups. TechCrunch articles offer valuable insights into various aspects of a startup, potentially including sustainability practices and other information relevant to valuation. Since Crunchbase originated as a database spun off from TechCrunch, the data formats exhibit a degree of similarity, facilitating the mapping of company names to their corresponding entries in Crunchbase.

This study restricts its scope to organisations that secure funding deals globally between 2010 and 2020. The initial dataset is downloaded on March 5, 2023, comprised 805,745 companies and 151,957 news articles. To focus on early-stage financing valuations, the analysis excludes funding rounds beyond Series J. Following this data filtering process, a final dataset of 772 unique companies with 1,112 funding rounds (seed to Series J) and 5,815 unique news articles is established.

Sustainability text

To explore the potential influence of sustainability-related information on startup valuation, this study incorporates data on two widely used standards[3] in sustainable finance: the United Nations' Sustainable Development Goals (SDGs) and the sustainability criteria established by the Sustainability Accounting Standards Board (SASB). Analysis of the news context surrounding each startup, employing the methodology explained in section 4.4.3, allows for the identification of connections between the information and the sustainability topics defined within these established frameworks.

- **UN Sustainable Development Goals (SDGs)**: This framework, established by the United Nations (2015); **?** outlines 17 interconnected goals aimed at achieving a sustainable future. The text description that explains the definition of each SDG's goal is used in the text similarity metrics.

- **Sustainability Accounting Standards Board (SASB)**: Developed by the International Sustainability Standards Board (ISSB), SASB provides industry-specific sustainability criteria for companies to report on 77 environmental, social, and governance issues tailored to their industry's impact (IFRS Foundation, 2022b). While SASB provides metrics for quantitative measurement of certain topics (e.g., percentage of food purchases meeting sustainability standards for the restaurant industry), such metrics are often unavailable for early-stage startups. Therefore, this study utilises the descriptive information provided for each topic within the SASB industry-specific documentation.

  Since the company dataset does not cover all industries listed by SASB, the companies are manually classified based on industry descriptions. This mapping results in a subset of 54 associated industries. Additionally, each industry-

[3] The data is collected on 3 July 2023.

specific disclosure topic is mapped to 26 broader topics encompassing five key dimensions: environment, social capital, human capital, business model and innovation, and leadership and governance (IFRS Foundation, 2022b).

Refinitiv

The study leverages a comprehensive ESG dataset from Refinitiv[4], a leading financial data provider. This data extends beyond traditional financial statements to include Environmental, Social, and Governance (ESG) information in the form of ESG pillar statements and ESG scores. Their ESG data is compiled from diverse sources, including company annual reports, websites, NGO websites, stock exchange filings, CSR reports, and new sources. By merging data from these varied sources, Refinitiv ensures a comprehensive picture of each company's ESG performance, ultimately enabling the calculation of a robust ESG score (Refinitiv, 2022). An example of ESG measurement is available in Appendix H.

Unlike publicly traded companies, early-stage private companies often lack publicly available information for the 2010-2020 timeframe. Furthermore, only two companies had third-party ESG scores, and none of the companies disclosed ESG statements within the specified timeframe. Due to these limitations, the study employs indicators for financial data disclosure and ESG scores to analyse the impact on data availability and company valuation.

4.5.1 Dependent variables

Startup valuations reflect their growth potential and market positioning, aiding venture capitalists (VCs) in pricing and determining funding potential. Two key valuation figures exist: pre-money and post-money valuations. This study explores the impact on a company's valuation prior to funding, focusing on pre-money valuation, calculated as the difference between post-money valuation and the funding amount.

The Kolmogorov-Smirnov test confirms a power-law distribution for both the raised funding amount and post-money valuation. Logarithmic transformation normalises the pre-money valuation distribution to a logarithmic scale, aligning with

[4] The data is collected on 1 August 2023.

the sampling approach used in prior research (Miloud et al., 2012; Gompers, 1995). Figure 4.2 visually depicts the log distribution of funding raised and pre-money valuation. It is important to note that the startups may raise funds multiple times within a year. To avoid duplication, only the most recent funding round per year for each startup is considered. The correlation coefficients between independent variables obtained from Crunchbase and pre-money valuation are presented in Appendix E.

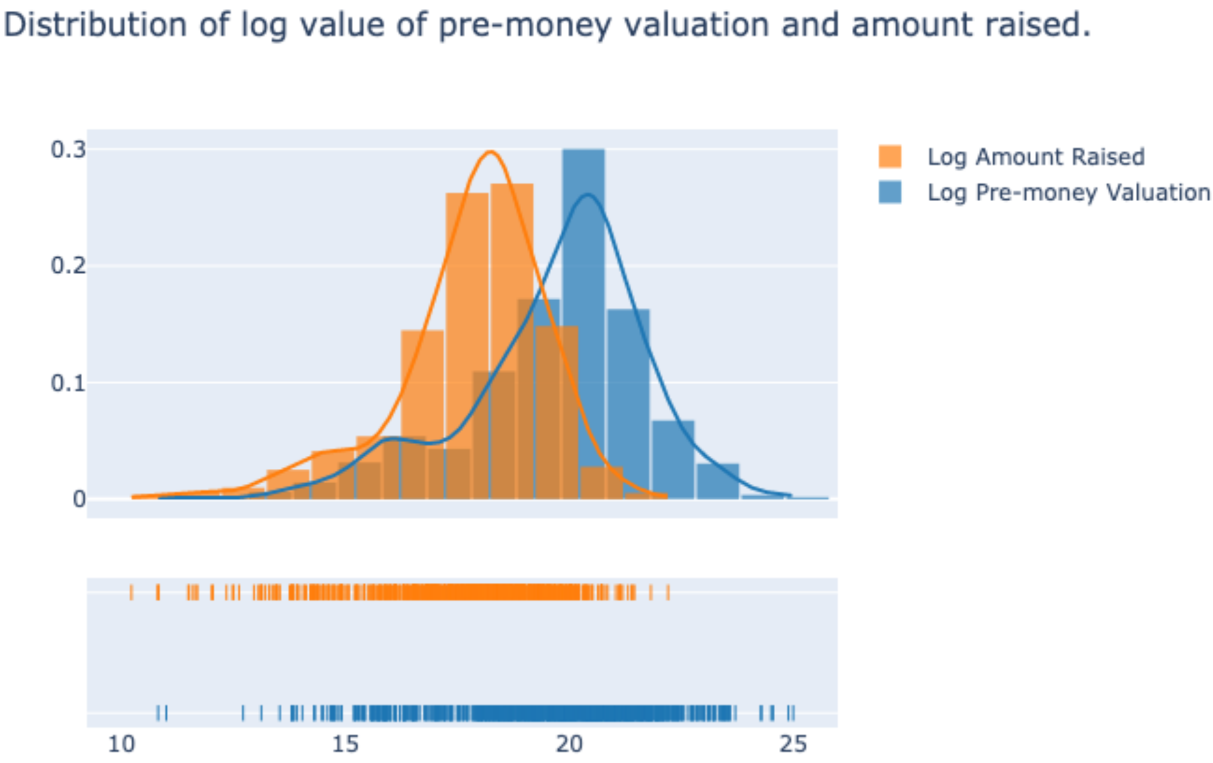

Figure 4.2: Distribution of funding amount raised and pre-money valuation between 2010 and 2020.

4.5.2 Statistics summary

Table 4.3 provides a detailed breakdown of all variables used in this study. The dataset comprises information on 1,112 funding rounds for 772 unique companies established and funded between 2010 and 2020. The initial data set is sourced from Crunchbase, as previously described.

Table 4.3: Summary statistics of Crunchbase features for predicting early-stage company valuation.

Variable Name	Count	Mean	Std	Min	Max
Organisation					
Country	772	24.350	10.048	0.000	35.000
Company Status	772	2.284	1.113	0.000	3.000
Has Facebook (Dummy)	772	0.911	0.285	0.000	1.000
Has Linkedin (Dummy)	772	0.970	0.170	0.000	1.000
Has Twitter (Dummy)	772	0.957	0.202	0.000	1.000
Employee Count	772	4.010	1.944	0.000	8.000
Company Founded Year	772	2013	2.276	2010	2020
Top Ten City (Dummy)	772	0.427	0.495	0.000	1.000
Founder & Co-Founder					
Founder Count	772	20.025	18.031	1.000	151.000
Has Bachelor	772	5.330	5.846	0.000	66.000
Has Master	772	2.197	2.902	0.000	27.000
Has MBA	772	2.053	2.524	0.000	18.000
Has PhD	772	0.558	1.156	0.000	10.000
Top 100 Education	772	3.373	4.003	0.000	35.000
Top 50 Education	772	2.845	3.528	0.000	31.000
Top 10 Education	772	1.826	2.526	0.000	19.000
STEM Education	772	3.312	3.928	0.000	36.000
Funding Round					
Investment Type	1112	3.676	2.484	0.000	11.000
Log Amount Raised	1112	17.809	1.711	10.224	22.205
Log Pre-money Valuation	1112	19.693	2.013	10.820	25.000
Deal Announced Year	1112	2017	1.976	2011	2020
Deal Age	1112	4.475	2.253	0.000	10.000
Number Funding Rounds	1112	4.499	3.065	0.000	23.000
Log Cum Amount Raised	1112	18.536	1.827	10.224	23.733

Table 4.3 continued from previous page

Variable Name	Count	Mean	Std	Min	Max
Investor					
Investor Count	1112	5.865	4.168	0.000	42.000
Top Institutional Investor (Dummy)	1112	0.180	0.384	0.000	1.000
Top Individual Investor (Dummy)	1112	0.003	0.052	0.000	1.000
Investment Age	1112	25.514	31.971	0.000	330.000
Accelerator Investor (Dummy)	1112	0.013	0.115	0.000	1.000
Angle Investor (Dummy)	1112	0.022	0.148	0.000	1.000
Different Geolocation (Dummy)	1112	0.550	0.498	0.000	1.000
Log Cum Invesment Amount	1112	7.732	10.536	0.000	26.061
Investor Experience	1112	96.745	131.617	0.000	759.000

Figure 4.3 illustrates the distribution of founding years. The years 2012 (16.32%), 2015 (14.37%), and 2011 (13.86%) witness the highest number of startup formations. The majority of startups (64.89%) are still operational, followed by those involved in M&A deals (15.15%), IPOs (13.73%), and those no longer operating (6.22%). Regarding social media presence, 91.06% of startups utilise Facebook, 97.02% use LinkedIn, and 95.72% have a Twitter presence.

Company size analysis reveals that 52.33% of startups fall into the medium-sized category with less than a thousand employees. Small companies with a team size of less than ten comprise 4.79% of the sample, while large corporations with over ten thousand employees represent 2.85%. Geographically, the majority of startups (65.29%) are based in the USA, followed by the United Kingdom (6.99%) and India (6.35%). Interestingly, only 42.75% of startups operate in the top ten cities globally. Product offerings show that software is the dominant category, accounting for 57.77% of startups. E-commerce and science and engineering companies comprise 19.30% and 17.49% of the sample, respectively.

The analysis of founder teams reveals an average size of 20 founders, which is higher than what might be typically observed in early-stage startups. On average, five

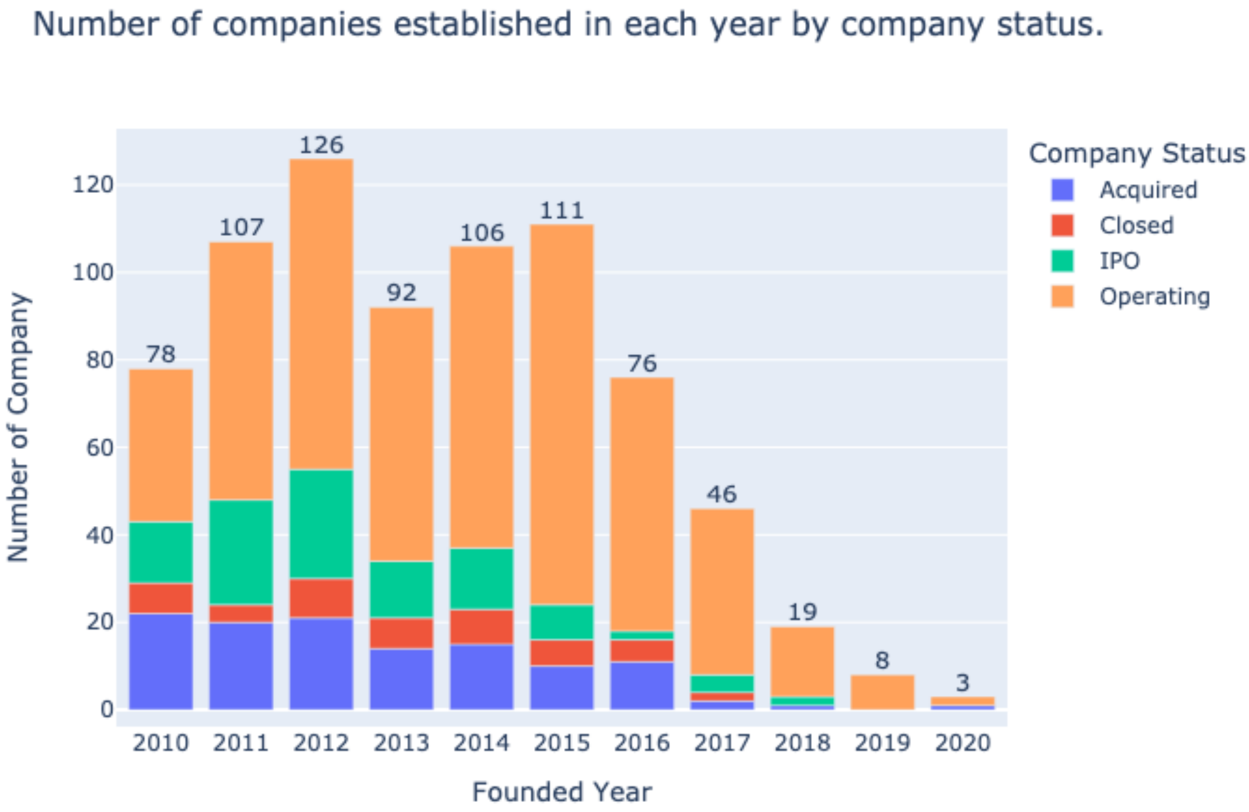

Figure 4.3: Distribution of companies by founding year and company status.

founders or co-founders hold bachelor's degrees, while at least two possess master's or MBA degrees. PhD qualifications are less common within the founder teams. Interestingly, at least one founder per team typically graduated from a top 10 globally ranked university. Furthermore, an average of two founders hold degrees from top 50 universities and three from top 100 universities. Notably, the average founder team of three members holds STEM degrees (Science, Technology, Engineering, and Mathematics).

Companies typically receive their first round of funding within a few years of establishment. However, many early-stage funding rounds (e.g., seed funding, Series A) often lack publicly available company valuation data. Consequently, the study primarily focuses on later-stage funding rounds, starting from Series C onwards, which typically occur a few years after company formation (Figure 4.4). For improved visualisation, samples are grouped into two categories: funding rounds before or equal to Series C, and those occurring after Series C.

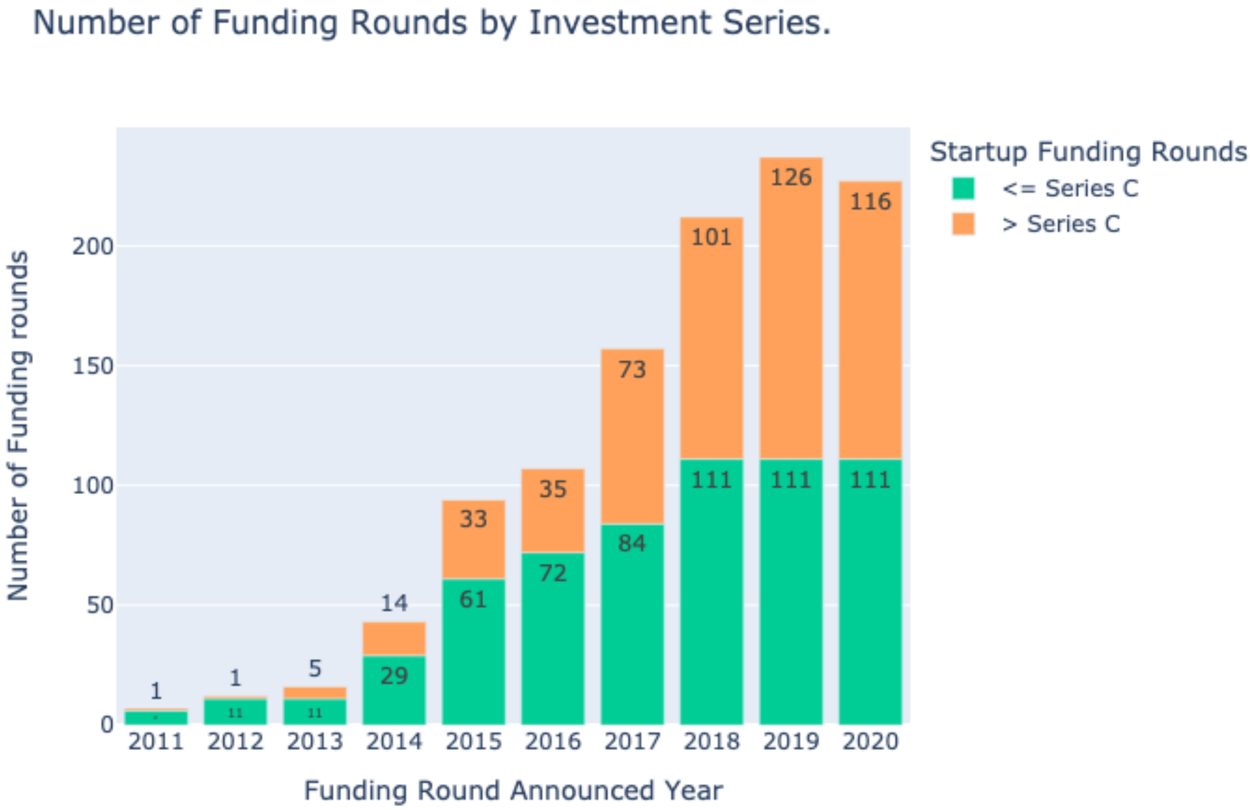

Figure 4.4: Distribution of funding rounds by deal announced year series stage.

The year 2019 witnesses the highest number of funding rounds (21.31%), followed by 2020 (20.4%) and 2018 (19.05%). On average, startups participate in approximately four funding rounds. The average time between company establishment and receiving funding is four years. The amount raised and valuation data are log-transformed for analysis, with average values of 17.81 and 19.69, respectively. The log-transformed average for the cumulative amount raised across all funding rounds is 18.54.

Each funding round typically attracts an average of five investors, with 18% being institutional investors and 0.3% individual investors. Investors possess an average of 25 years of experience and participate in an average of 96 funding rounds by the time of their current investment. Accelerators or incubators play a role in 1.3% of the funding rounds, while 2.2% of startups previously receive funding from angel investors. Moreover, in 55% of cases, the investor location differed from the startup location. The average log-transformed cumulative investment amount raised by investors is 7.732.

The average number of news articles associated with each company in the dataset is eight. However, this number varies considerably across companies. For instance, companies like Lyft and Pinterest, which are established near the beginning of the study period, have a significantly higher number of news articles (over one hundred) due to greater historical public exposure. The distribution of news articles by publication year reveals that 2019 has the highest concentration (20.91%), followed by 2020 (16.53%) and 2018 (13.91%).

4.6 Results

4.6.1 Valuation prediction

Aligned with the study's objective of predicting pre-money valuation for early-stage companies, Appendix I presents the significant results of predictions based on variables obtained from the Crunchbase dataset and company news from TechCrunch.

Crunchbase

Analysing the Crunchbase data (Table 4.4), which includes variables relevant to founders, companies, and investors participating in funding rounds, linear regression yields the highest R^2 of 0.747 with a low MSE value of 0.617. Gradient boosting followed closely with a second-lowest MSE value of 0.626 and an R^2 of 0.743. Interestingly, the simpler linear regression model outperforms ensemble methods like random forest and complex deep learning models.

Table 4.4: The performance metrics exhibited by various machine learning models employed for startup valuation prediction. The models include linear regression (LR), random forest (RF), gradient boosting (GB), and deep learning (DL). Only the results with the highest R-squared grouped by the combination of text embedding and supervised learning are presented.

Model	R-Squared	MSE	RMSE	MAE
Crunchbase				
LR	**0.747**	**0.617**	**0.785**	**0.546**
RF	0.691	0.752	0.867	0.637
GB	0.743	0.626	0.791	0.566
DL	0.541	1.118	1.057	0.802
TechCrunch				
Doc2Vec + GB	-0.039	2.589	1.609	1.252
BERT + RF	**0.007**	**2.475**	**1.573**	**1.219**
FinBERT + RF	-0.017	2.536	1.592	1.235
ESG-BERT + RF	-0.022	2.548	1.596	1.243
Crunchbase & TechCrunch				
Doc2Vec + GB	**0.761**	**0.595**	**0.771**	**0.614**
BERT + LR	0.727	0.681	0.825	0.616
FinBERT + DL	0.721	0.696	0.835	0.667
ESG-BERT + LR	0.676	0.807	0.899	0.674

Despite a significant R^2, the linear regression model suggests a potential violation of the linearity assumption. Figure 4.5 depicts a scatter plot of the model's residuals for pre-money valuation and various Crunchbase variables. The observed non-linear pattern in the residuals indicates that a linear model may be inadequate to capture the relationship between these variables. This aligns with the findings by Ang et al. (2022), who observe non-linear relationships between determinants and predicted values. Further investigation confirms high collinearity among independent variables and non-normal distribution of residuals.

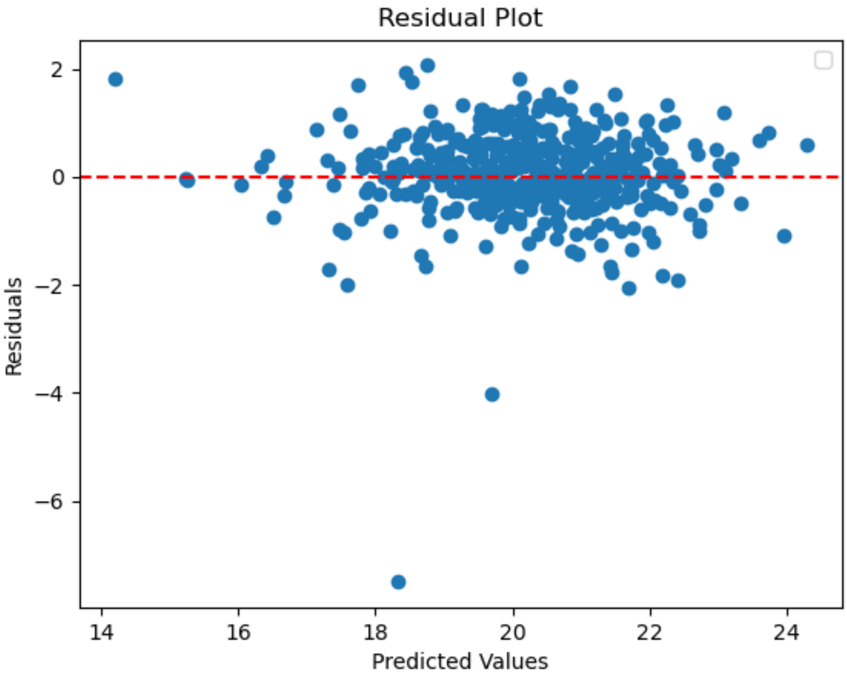

Figure 4.5: Residual Plot of Crunchbase variables and Pre-Money Valuation

Therefore, given the non-linearity, gradient boosting emerges as the second-best performing model with a comparable R^2 of 0.743 and a slightly lower MSE of 0.626. These results suggest that data from Crunchbase alone holds promise for predicting the pre-money valuation of young companies.

TechCrunch and Text Embedding Models

Analysis of news data from TechCrunch, however, suggests limited predictive power for valuation. As shown in Table 4.4, most models incorporating solely news data yield negative R^2 scores, indicating a poor model fit. The only exception involves text embeddings generated by the BERT model in conjunction with a random forest model for prediction. Nevertheless, this approach results in a low R^2 of 0.007 and a high MSE of 2.475. These findings suggest that sentence embeddings of news titles alone are insufficient for determining company valuation.

Further investigation into the impact of different embedding models, including Doc2Vec, BERT, FinBERT, and ESG-BERT (Table 4.5), yields similar results. Despite the use of domain-specific models like FinBERT and ESG-BERT, pre-trained on business and financial corpora, the models still generate negative R^2 values and high MSE scores, indicating insufficient predictive power for valuation. Notably, deep learning models perform even worse than ensemble algorithms like a random forest.

Table 4.5: The performance metrics achieved by various text embedding models employed for startup valuation prediction. The models include Doc2Vec, BERT, FinBERT, and ESG-BERT.

Model	R-Squared	MSE	RMSE	MAE
TechCrunch				
Doc2Vec + LR	-0.398	3.484	1.867	1.467
Doc2Vec + RF	-0.157	2.884	1.698	1.325
Doc2Vec + GB	**-0.039**	**2.589**	**1.609**	**1.252**
Doc2Vec + DL	-0.513	3.771	1.942	1.535
BERT + LR	-0.486	3.704	1.924	1.505
BERT + RF	**0.007**	**2.475**	**1.573**	**1.219**
BERT + GB	-0.021	2.545	1.595	1.239
BERT + DL	-0.235	3.078	1.755	1.369
FinBERT + LR	-0.352	3.369	1.835	1.455
FinBERT + RF	**-0.017**	**2.536**	**1.592**	**1.235**
FinBERT + GB	-0.024	2.553	1.598	1.241
FinBERT + DL	-0.178	2.937	1.714	1.350
ESG-BERT + LR	-0.431	3.567	1.889	1.485
ESG-BERT + RF	**-0.022**	**2.548**	**1.596**	**1.243**
ESG-BERT + GB	-0.025	2.555	1.599	1.242
ESG-BERT + DL	-0.423	3.547	1.883	1.467

Crunchbase and TechCrunch

Combining Crunchbase data with news data from TechCrunch (Table 4.6) yields promising results. The Doc2Vec model for embedding news titles performs particularly well, with gradient boosting achieving the highest R^2 (0.761) and lowest MSE (0.595) among all models. The results demonstrate the potential of incorporating unstructured text information to improve startup valuation prediction, as evidenced by the outperformance of our approach over the gradient boosting model using only Crunchbase data (0.626 in Table 4.4).

Table 4.6: The performance metrics achieved by various machine learning models and text embedding techniques, using a combination of the Crunchbase dataset and company news from TechCrunch.

Model	R-Squared	MSE	RMSE	MAE
Crunchbase & TechCrunch				
Doc2Vec + LR	0.754	0.613	0.783	0.567
Doc2Vec + RF	0.736	0.658	0.811	0.646
Doc2Vec + GB	**0.761**	**0.595**	**0.771**	**0.614**
Doc2Vec + DL	0.725	0.687	0.829	0.63
BERT + LR	**0.727**	**0.681**	**0.825**	**0.616**
BERT + RF	0.432	1.416	1.190	0.981
BERT + GB	0.219	1.946	1.395	1.100
BERT + DL	0.667	0.829	0.911	0.711
FinBERT + LR	0.706	0.732	0.855	0.637
FinBERT + RF	0.434	1.410	1.187	0.983
FinBERT + GB	0.219	1.947	1.395	1.101
FinBERT + DL	**0.721**	**0.696**	**0.835**	**0.667**
ESG-BERT + LR	**0.676**	**0.807**	**0.899**	**0.674**
ESG-BERT + RF	0.427	1.428	1.195	0.993
ESG-BERT + GB	0.219	1.946	1.395	1.100
ESG-BERT + DL	0.640	0.897	0.947	0.760

The general-purpose BERT model yields a mixed performance, with an average MSE of 1.218. Only the BERT model with linear regression outperforms other BERT models. However, given the non-linear relationships observed in the data, deep learning models with BERT embeddings are expected to capture these complexities. While the BERT model with deep learning achieves an MSE of 0.829 (still considered slightly high), the FinBERT model outperformes general-purpose BERT models with a lower MSE of 0.696. The overall average MSE for FinBERT integration is 1.196.

Interestingly, the ESG-BERT model, designed for the sustainability domain, underperformed the general-purpose BERT model, exhibiting both higher MSE and lower R^2 values. It is worth noting that the average MSE score for ESG-BERT across regression models (1.269) is the highest among all embedding models used in this study.

Crunchbase and Sustainability Similarity

While transforming text data into numerical features for regression analysis can be informative, this study employs text embedding models to convert news and sustainability-related text into vectors before calculating similarity. Table 4.7 demonstrates the impact of integrating sustainability sources from the UN SDGs, SASB reports, and Refinitiv statements. Among the models incorporating these sources, the random forest model emerges as the most effective. It outperforms the simpler linear regression model that relies solely on Crunchbase data with Doc2Vec embeddings and gradient boosting.

Table 4.7: The performance metrics achieved by various machine learning models and text embedding techniques, integrating text similarity with UN Sustainable Development Goals (SDGs) and Sustainability Accounting Standards Board (SASB) criteria, along with ESG disclosure data obtained from Refinitiv.

Model	R-Squared	MSE	RMSE	MAE
Crunchbase				
LR	**0.747**	**0.617**	**0.785**	**0.546**
RF	0.691	0.752	0.867	0.637
GB	0.743	0.626	0.791	0.566
DL	0.541	1.118	1.057	0.802
Crunchbase & UN SDG				
LR	0.753	0.602	0.776	0.542
RF	**0.775**	**0.549**	**0.741**	**0.515**
GB	0.756	0.596	0.772	0.545
DL	0.550	1.097	1.048	0.788
Crunchbase & SASB General Issue				
LR	0.690	0.755	0.869	0.631
RF	**0.785**	**0.523**	**0.723**	**0.503**
GB	0.770	0.561	0.749	0.524
DL	0.456	1.327	1.152	0.898
Crunchbase & SASB Dimension				
LR	0.746	0.618	0.786	0.547
RF	**0.780**	**0.536**	**0.732**	**0.507**
GB	0.715	0.695	0.833	0.613
DL	0.532	1.140	1.068	0.802
Crunchbase & Refinitiv				
LR	0.747	0.617	0.786	0.547
RF	**0.772**	**0.556**	**0.745**	**0.519**
GB	0.723	0.676	0.822	0.587
DL	0.543	1.115	1.056	0.818

Leveraging the similarity between news content and the 26 general issues highlighted by SASB, the model achieves the lowest MSE score (0.523) and a high R^2 (0.785). The model using aggregated data from the five SASB dimensions followed closely with an R^2 of 0.780. Notably, the average MSE difference between models using SASB general issues and those using SASB dimensions is only 5.93%.

Integrating similarity scores between the descriptions of the 17 UN SDGs and news articles also yields positive results, with an MSE of 0.549 and an R^2 of 0.775. Adding financial disclosure data and ESG scores further improves the model, with the random forest model achieving an R^2 of 0.772 and an MSE of 0.556. This finding is particularly noteworthy given the limited availability of ESG data within the study timeframe. These results suggest that incorporating the sustainability context, even with limited ESG data, can significantly enhance the model's performance compared to relying solely on standard financial and non-financial variables from Crunchbase.

This study compares the performance of various machine learning models for startup valuation prediction. Ensemble regression models, particularly random forest, yield generally good results compared to other regression models. Unfortunately, none of the deep learning models achieve superior performance to the baseline regression models in contrast to the hypothesis H_2. This suggests that the limited size and complexity of the startup dataset might hinder the ability of deep learning models to effectively capture the valuation process.

Furthermore, while text embedding models pre-trained on ESG and financial corpus (ESG-BERT and FinBERT) demonstrate promising results, they do not outperform simpler embedding models like Doc2Vec and general-purpose BERT. This finding suggests that, in this context, simpler embedding techniques may be sufficient for capturing relevant information from text data for valuation prediction, contradicting the hypothesis H_3. This observation warrants further investigation into the optimal balance between domain-specific and general-purpose text embedding for startup valuation tasks.

4.6.2 Feature importance

Tables 4.4 and 4.7 demonstrate the value of integrating sustainability text data with Crunchbase data to improve early-stage company valuation prediction. To gain a deeper understanding of how individual variables contribute to the model's performance, the analysis of feature importance is conducted. This analysis focuses on the top-performing model from Table 4.7, a random forest model that utilises Crunchbase data and similarity scores with the 26 SASB general issues. This model achieves the lowest MSE score (0.523) among all models investigated.

Figure 4.6 illustrates the top ten determinants with the highest feature importance values in the random forest model. The log value of the cumulative amount of funding raised by the company exhibits the highest relative importance (0.858), followed by the log value of the amount raised in a single round (0.042). This substantial difference in importance scores suggests that the cumulative funding history holds greater weight in predicting valuation compared to the size of individual funding rounds. This finding aligns with Ang et al. (2022); Garkavenko et al. (2021), who highlighted the use of historical funding rounds by venture capitalists as a crucial reference point for determining investment decisions and valuation. The remaining determinants in the analysis yield relatively low feature importance scores.

Building upon prior research that emphasises the critical role of founders (Macmillan et al., 1985; Miloud et al., 2012; Hsu, 2007; Damodaran, 2009), the result indicates that the educational background (bachelor's degree) and the number of founders significantly impact the prediction of early-stage company valuation. Additionally, the model highlights the relevance of specific sustainability factors. In the sustainability context, three SASB general issues contribute to valuation: (1) encouraging equal opportunities for workforce satisfaction, (2) fostering fair competition and market practices that protect customer welfare, and (3) adopting energy-efficient practices and renewable energy in their operations. This suggests that investors may actively consider a company's approach to these specific sustainability aspects when making valuation decisions. Finally, the number of funding rounds and company size also emerge as significant factors influencing valuation.

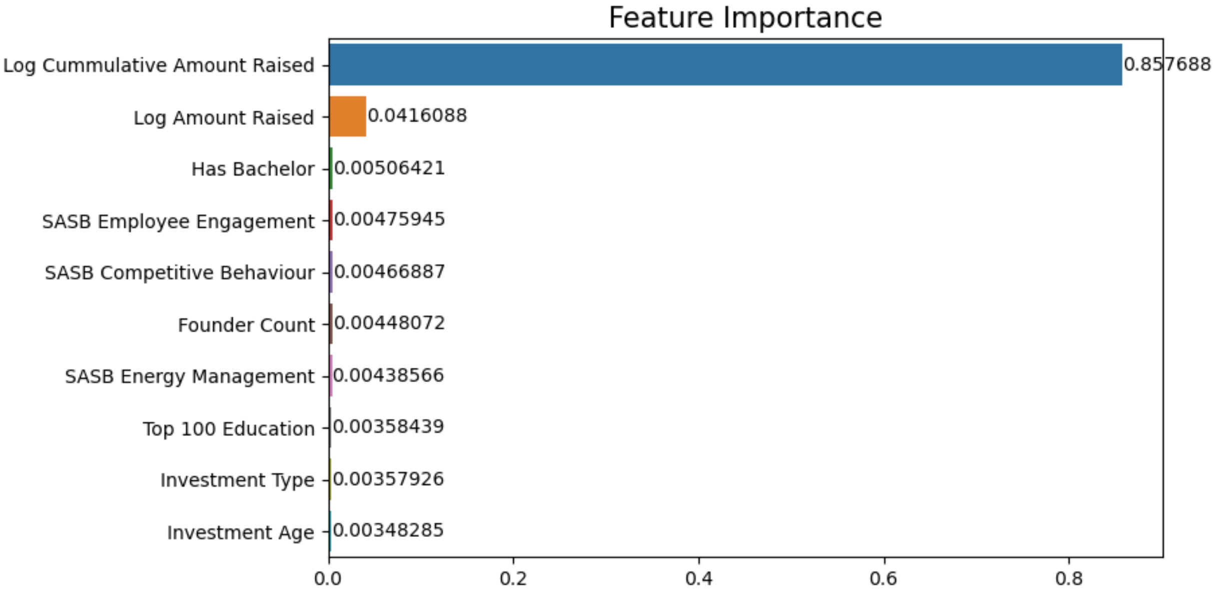

Figure 4.6: Feature importance of the Random Forest model gives n the variables from Crunchbase and SASB general issue text similarity.

Feature Importance of Sustainability Indicators

Beyond founder, startup, and market characteristics influencing valuation, this study investigates the relationship between news titles and sustainability-related text derived from the UN SDGs and SASB. Samples are further categorised based on funding rounds: pre-Series C (including Series C) and post-Series C.

Figure 4.7 presents the feature importance of text similarity to UN Sustainable Development Goals (SDGs), revealing significant differences between the two sample groups. At the early-stages, it focuses on “Gender Equality” for founding the team. For companies beyond Series C funding, “Sustainable Cities and Communities” exhibits the highest feature importance. This can indicate how startups positively impact communities, encouraging sustainable economic, environmental, and social growth. These findings suggest that investor priorities regarding sustainability may shift as companies mature and progress through funding stages.

Similar to the findings for UN SDGs, the relative importance of text similarity scores with the 26 SASB general issues and dimensions reveals distinct patterns across funding stages. Figures 4.8 and 4.9 illustrate a strong focus on human capital in early funding rounds. This is reflected in the high feature importance of “Employee

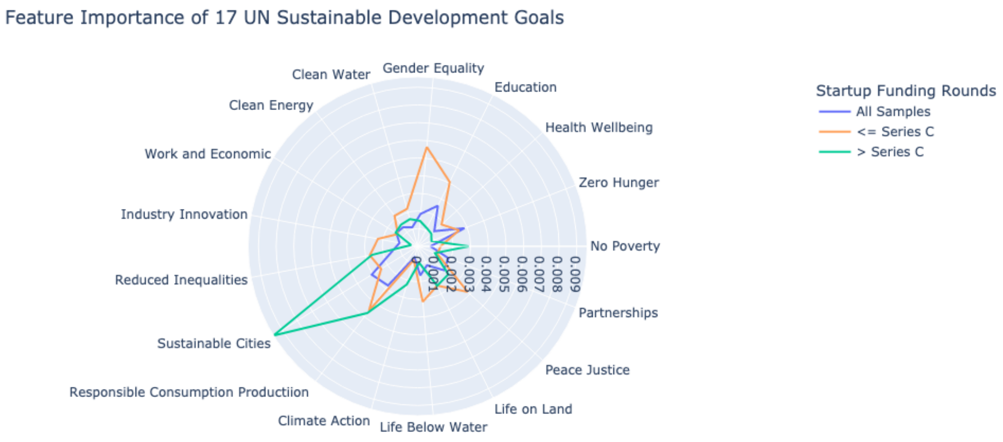

Figure 4.7: Feature Importance of UN SDGs in Startup Valuation.

Engagement, Diversity and Inclusion," and the broader "Human Capital" dimension.

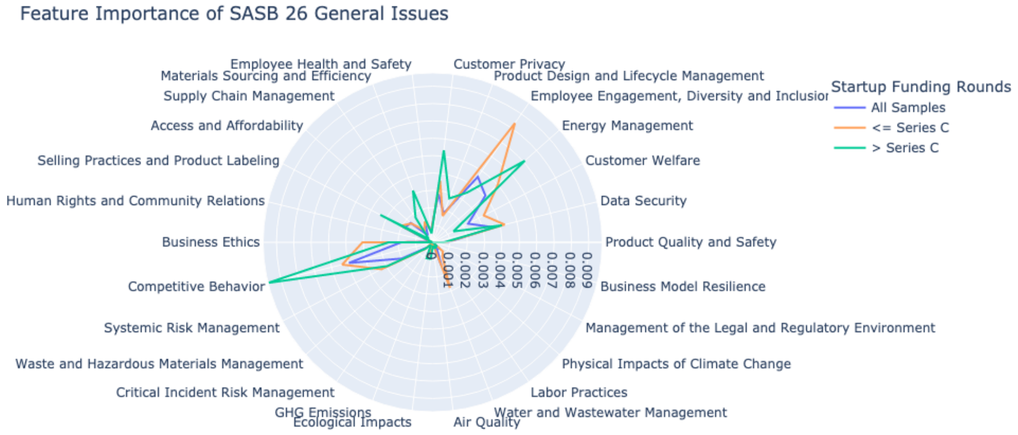

Figure 4.8: Feature Importance of 26 SASB general issues in Startup Valuation.

For companies that secure funding beyond Series C, the relative importance shifts. "Competitive Behavior" emerges as the most significant factor based on SASB general issue similarity scores. Additionally, within the high-level SASB dimensions, the focus initially placed on "Leadership and Governance" diminishes, while other dimensions contribute more equally to valuation predictions.

This study highlights the influence of specific sustainability topics on startup valuation across funding stages. Notably, "Human Capital" and "Employee Engage-

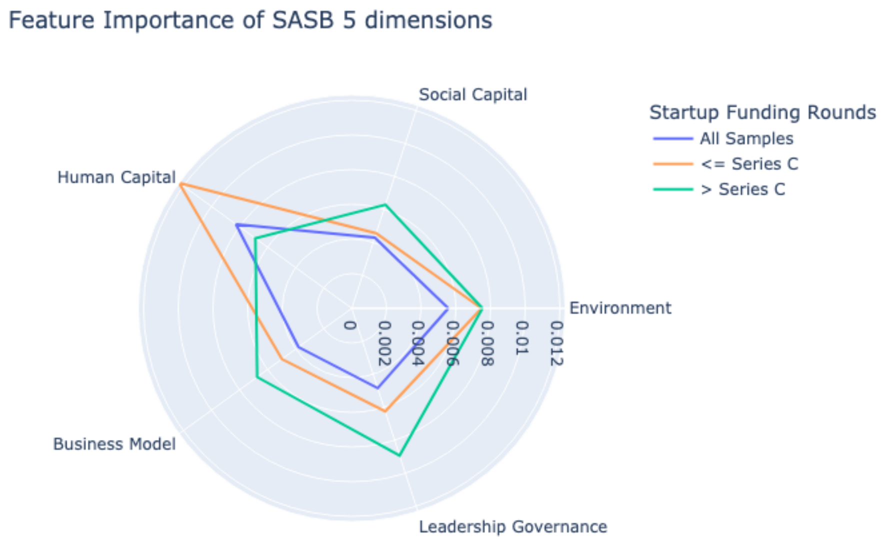

Figure 4.9: Feature Importance of 5 SASB dimensions in Startup Valuation.

ment" emerge as highly impactful factors for early-stage companies (Figures 4.8, 4.9). This suggests that investors in early funding rounds may place a high value on a company's approach to talent management and employee well-being. However, investor priorities appear to shift for companies that secure funding beyond Series C. Here, "Competitive Behavior" and "Leadership and Governance" within the SASB framework take on greater importance, reflecting a focus on long-term strategic positioning and sustainable business practices.

While the relative importance of sustainability factors remains lower than core valuation drivers like historical funding and founder characteristics, their integration demonstrably improves model performance, confirming the hypothesis H_1. Compared to the baseline model, the inclusion of sustainability data yields a significant 16.45% reduction in MSE score. This finding suggests that considering sustainability aspects alongside traditional financial metrics can enhance the accuracy of early-stage company valuation models.

4.7 Discussion

This study leverages machine learning and natural language processing (NLP) techniques to exploit unstructured text data. This approach addresses the challenges associated with private market valuation and the integration of sustainability factors into investment decisions. The proposed model offers the following theoretical and practical implications:

4.7.1 Theoretical implication

Traditional valuation methods often rely on financial statements to predict discounted cash flow (Williams, 1938). However, such data can be scarce, particularly in the private equity and venture capital landscape. This study proposes a machine learning-based valuation approach that utilises unstructured text data as an alternative data source to overcome these limitations. While text embedding models effectively convert text data into vectors suitable for machine learning analysis, these vectors do not directly correspond to financial attributes used in traditional regression models.

The proposed model addresses this by integrating text embedding models to capture the similarity of sustainability text data with established startup, founder, and market characteristics explored in prior literature. The feature importance analysis reveals valuable insights. The raised funding amount emerges as a key factor, reflecting the well-established practice of venture capitalists using historical funding rounds as a reference point for future valuations. This aligns with the findings of Ang et al. (2022); Garkavenko et al. (2021). The results also demonstrate a notable shift in valuation attributes observed before and after Series C funding within the sustainability context, as indicated by the similarity scores. This suggests that venture capitalists' interests evolve from primarily focusing on human capital and talent management, which serve as foundational strengths for early startup success, towards assessing long-term strategic positioning and sustainable business practices. This change reflects a refined understanding of what drives sustained growth and competitive advantage in the market.

This study highlights the potential of incorporating sustainability factors into early-stage company valuation models. By investigating this under-explored area within academic research, the study contributes to the promotion of economically, environmentally, and socially responsible development within private capital investment (Cumming et al., 2022). As investor interest in sustainable investment grows, sustainability considerations are likely to evolve into significant valuation attributes. This aligns with research in public equities, which demonstrates the impact of sustainability factors and news articles on asset returns (Guo et al., 2020; Schmidt, 2019).

Future research should expand beyond startup news and general sustainability scores to investigate other relevant sustainability features. For example, the presence of gender bias at both management and employee levels can significantly impact a startup's ability to raise funding and influence its valuations (Kanze et al., 2018). Alternatively, future studies can assess and measure startup competitiveness by considering attributes such as innovation, intellectual property, and investment in Research and Development (R&D) (Silva Júnior et al., 2022). These attributes can then be translated into quantitative features and incorporated into valuation prediction models.

4.7.2 Practical implication

Machine learning offers significant potential as a data-driven tool for venture capitalists. By facilitating efficient data processing and execution of due diligence tasks, machine learning can help overcome information asymmetries inherent to the illiquid private capital market. This study demonstrates the potential of leveraging unstructured text data, such as news articles, as a complementary data source for screening investment opportunities and company valuation.

Integrating sustainability-context text analysis can serve as an additional signal to support two key investment strategies: (1) incorporating sustainability into existing funds or (2) establishing dedicated sustainability-focused funds (Lin, 2022). This integration of sustainability factors has the potential to strengthen investor and limited partner confidence in these emerging investment trends. Furthermore, the research encourages practitioners in the private capital market to acknowledge the critical role of sustainability, encompassing not only short-term financial gains but also long-term environmental and social impact.

However, a significant information gap exists in both financial and sustainability-related data availability for private companies. Despite the presence of established data providers like Refinitiv, generating comprehensive ESG statements within the private capital market remains challenging. As Figure 4.10 illustrates, the majority of ESG disclosures are concentrated among publicly traded companies that complete initial public offerings (IPOs). Similarly, Figure 4.11 demonstrates that the United States possesses the highest rate of data disclosure, followed by the United Kingdom, while most other countries lag significantly behind. Even with the absence of formal sustainability reports from startups, news data pertaining to these ventures presents a promising alternative dataset for bridging this information gap. Additionally, policymakers and regulators can play a crucial role in encouraging sustainability disclosures within the private capital market.

Furthermore, this research aligns with Lin (2022) in advocating for the development of clear sustainability definitions and standards within specific regions, markets, and regulatory frameworks. Such measures can help minimise greenwashing practices. Additionally, as suggested by Bianchini and Croce (2022), governments can play a crucial role by enacting policies that create long-term demand for sustainable products and provide financial grants to support startups in differentiating themselves for long-term fundraising success. Regulatory bodies can also play a part, as exemplified by the European Union's SFDR (Sustainable Finance Disclosure Regulation). This regulation mandates that all financial market participants, including venture capital firms, comply with sustainability disclosure requirements. This includes highlighting their risk policies and demonstrating how sustainability risks are integrated into decision-making processes, with potential consequences reflected in remuneration plans (Roure, 2024; European Comission, 2022).

Beyond the investor perspective, startups can also develop innovative solutions in response to growing sustainability demands. Industries like cleantech offer promising opportunities for such innovation. However, fostering such innovation requires a supportive ecosystem within the private market. Financial institutions, academic

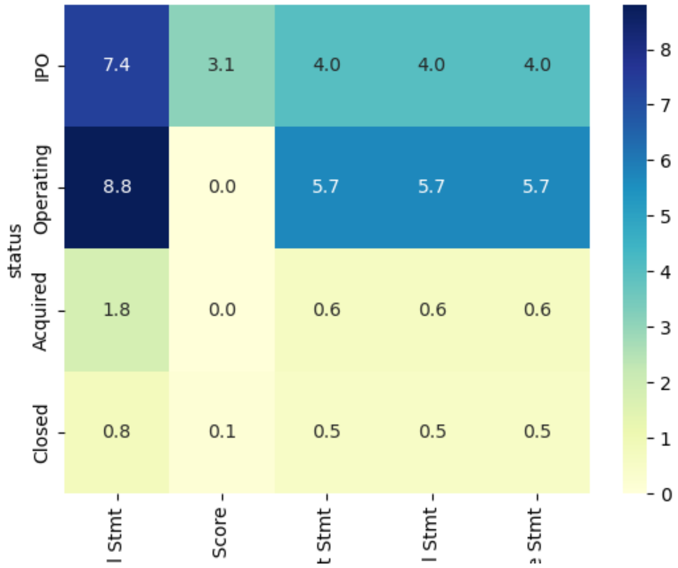

Figure 4.10: Completeness of ESG Data Disclosure available in Refinitiv grouped by the company status.

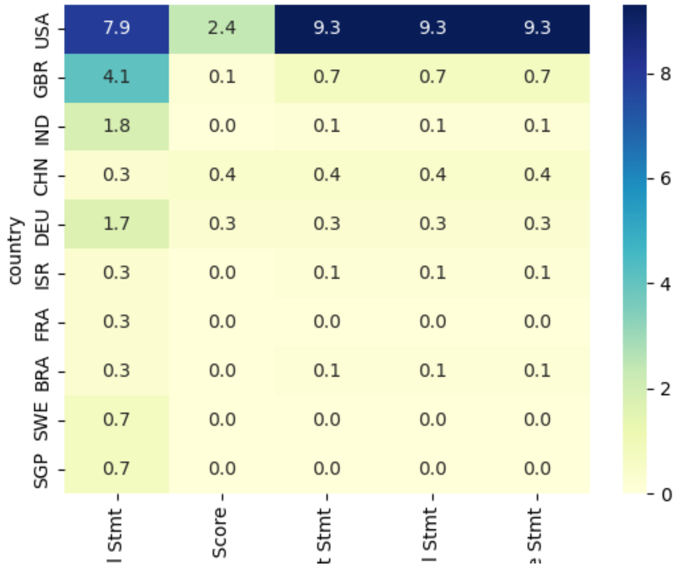

Figure 4.11: Completeness of ESG Data Disclosure available in Refinitiv grouped by the country.

institutions, governments, and other stakeholders like accelerators and incubators form a critical ecosystem that empowers early-stage companies to create value and respond effectively to sustainability demands from both markets and regulators.

While the importance of sustainability is undeniable, investors must also consider other relevant factors not explicitly addressed in this study. Market conditions, including interest rates, can profoundly affect the funding supply, thereby influencing startup valuations and competition among VCs (Hsu, 2007). This occurs as these conditions significantly impact limited partners' (LPs) ability to contribute capital and, consequently, venture capitalists' capacity to invest in new ventures. These additional considerations require careful analysis alongside sustainability aspects for informed investment decisions.

4.8 Conclusion

In response to the growing demand for sustainable investment within the private equity and venture capital landscape, this study proposes a novel model that explores the potential of integrating sustainability factors and unstructurecd text data into valuation models for early-stage companies. The findings reveal that sustainability considerations are increasingly influencing investor decisions and company valuations, particularly for startups in their early funding rounds. Companies that prioritise and effectively communicate their sustainability efforts stand to benefit from higher valuations.

Unstructured text data, such as news articles, offers a valuable alternative data source to address the information scarcity inherent in private capital markets. By leveraging machine learning and text embedding models, this data can be incorporated into valuation models to capture relevant information that may not be readily available in traditional financial statements. Interestingly, the study found that domain-specific embedding models do not significantly enhance model performance. However, the similarity between startup news content and sustainability text derived from frameworks like the SASB and UN SDGs significantly improves valuation prediction, achieving a 16.45% improvement over the baseline regression.

Feature importance analysis reveals that traditional factors such as founder characteristics continue to play a significant role in company valuation. However, the funding amount raised by the company emerges as a particularly strong indicator, reflecting the practice of venture capitalists using historical funding rounds as a reference point for future company valuations. The analysis significantly reveals a strategic shift in sustainability priorities, indicating a transition from an emphasis on team and talent development towards strategic positioning that impacts society in the long term.

A significant information gap persists regarding both financial and sustainability-related data for private companies. Regulatory bodies and policymakers can play a critical role in bridging this gap by encouraging sustainability disclosures within the private capital market. While sustainability is gaining traction as an investment factor, investors must remain mindful of other relevant considerations. Other relevant factors can significantly impact the ability of investors to secure funding (from limited partners) and, consequently, their capacity to invest in startups. These additional considerations require careful analysis alongside sustainability aspects for informed investment decisions.

This study lays the groundwork for further exploration in several key areas. Firstly, incorporating additional sources of unstructured text data beyond TechCrunch can potentially enrich the model's understanding of company performance. Analysing sentiment within news articles can provide valuable insights into market perception, while examining the impact of ESG-related news controversies on company value presents another intriguing research avenue. Techniques like data augmentation, as employed by Nugent et al. (2021), can be leveraged to address information scarcity. Furthermore, investigating the relationships between unstructured text and various sustainability factors can provide valuable insights. This may involve extending the analysis to predict future company statuses, such as IPO prospects, receipt of additional funding, or closure. Finally, integrating advanced large language models (LLMs) like GPT-4 and Llama 2 into the existing framework represents a potential avenue for future research. These models may offer superior capabilities for extracting and processing information from unstructured text data, potentially leading to further improvements in model performance.

Chapter 5

Identifying High-Growth Startups: A Reinforcement Learning Approach for Venture Capital

5.1 Introduction

Venture capital (VC) transcends mere financial deployment anticipating high returns. It actively contributes to a company's value creation by fostering sustainable growth, benefiting stakeholders, and enhancing future value. Given the vast and dynamic landscape of operational startups driving economic progress, VC investors face growing challenges in selecting promising ventures. The initial stage of the VC process involves screening and selecting startups that align with the firm's investment strategy. This strategy outlines targeted sectors and preferred startup stages, allowing the VC team to leverage their industry knowledge and experience to drive value creation within the portfolio companies. This ongoing process also includes continuous support and monitoring of these portfolio companies, while ensuring alignment with the interests of the VC firm's limited partners.

Following the screening process, VC firms perform thorough due diligence to assess various aspects, including the founders, team, product/service offerings, and market potential, to determine if the company holds the potential to generate future financial returns. Traditionally, this approach relies heavily on direct interaction with founders, secondary market research, or intuition-based decision-making. This subjectivity can lead to inconsistencies and biases in the selection process, potentially causing VC firms to miss out on promising investment opportunities. Nonetheless, the emergence of big data and artificial intelligence (AI) spurs a paradigm shift towards a data-driven approach within VC firms. Several academic studies explore the application of machine learning models, particularly supervised learning algorithms, to predict company exit events and valuation using training datasets (Bhat and Zaelit, 2011; Arroyo et al., 2019). These supervised models excel at identifying patterns in existing data, allowing them to make predictions based on past performance.

While widely adopted in financial applications, supervised learning models exhibit limitations within the dynamic environment of VC. These models rely heavily on large, labelled training datasets, which presents a significant hurdle in the VC context due to the scarcity of historical data. To address these limitations, reinforcement learning (RL) presents a promising alternative that receives limited exploration within the private capital space. RL employs an agent that interacts with its environment and selects actions to maximise long-term rewards. Recent applications in public equities demonstrate the potential of RL to simulate investment decision-making processes (Azhikodan et al., 2019; Charpentier et al., 2023; Deng et al., 2017; Bühler et al., 2018; Moody and Saffell, 2001; Spooner et al., 2018). However, its application as a recommendation system within the financial domain remains largely unexplored (Afsar et al., 2022). This capability offers a compelling solution for VC by enabling the implementation of RL models as recommendation systems for promising startups. The RL model's ability to continuously adapt its selection criteria based on past successes has the potential to outperform traditional methods employed by VC firms. This study aims to bridge the existing research gap by investigating the effectiveness of RL in identifying optimal investment candidates for VC funds.

Venture capital portfolio selection is a complex task marked by inherent uncertainty and information asymmetry (Denis, 2004). This research introduces the Venture Capital Reinforcement Learning Recommender System (VC-RLRS), a novel framework designed to address the limitations of traditional VC investment strategies. Unlike its applications in e-commerce, the potential of reinforcement learning recommender systems remains largely unexplored within financial applications, particularly in illiquid markets like venture capital and private equity. The groundbreaking approach, built upon a Q-learning model, offers a significant advance by incorporating state representations and reward functions uniquely tailored to the VC domain.

Leveraging reinforcement learning's strengths, the VC-RLRS effectively learns optimal investment strategies through interaction with simulated investment environments. The model demonstrates the capability to recommend startups with high growth potential while explicitly considering crucial factors like exit opportunities and portfolio diversification, conceptually extending the application of recommender systems to complex financial ecosystems. It showcases its potential to enhance investment decision-making for both generalist and specialist strategies, with successful demonstrations across specific industries like FinTech, Healthcare, and Information Technology. Furthermore, this study makes a methodological contribution by integrating deep learning techniques within a hybrid model to assess its performance against a baseline Q-learning approach, simultaneously highlighting areas for future scalability improvements. Overall, this research presents an original design and evaluation of reinforcement learning components specifically for the VC context, significantly contributing to the field while outlining promising future directions for practical application by VC practitioners.

The remainder of this chapter is structured as follows. Section 2 provides background on machine learning and reinforcement learning models in financial applications. Sections 3, 4, and 5 introduce the fundamentals of reinforcement learning, along with the experiment design choices and dataset descriptions. Section 6 presents the study's results, followed by a discussion of the findings, implications, and limitations in Section 7. Finally, Section 8 concludes the research and outlines future work directions.

5.2 Literature review

5.2.1 Evaluating investment opportunities in venture capital

Early-stage companies play a vital role in the global economy by driving innovation, fostering job creation, and stimulating financial flows within the ecosystem encompassing customers, suppliers, and investors. Investment in these companies comes from various sources, including angel investors, venture capitalists (VCs), and individuals participating in collective forms like crowdfunding. However, startup investment is inherently risky, with a significant 20% failure rate within the first year (U.S. Bureau of Labor Statistics, 2022). VCs hold a distinct advantage among early-stage investors due to their superior access to information. While their collective experience enables them to identify promising opportunities, constructing well-diversified portfolios remains a complex challenge, hindered by inefficient processes and information asymmetries.

Identifying high-potential startups and cultivating them into successful portfolio companies remains a core challenge for venture capital (VC) firms. VCs must navigate a complex landscape of signals and criteria to assess risks and opportunities, ultimately aiming to select outstanding firms with the potential for extreme future returns. This challenge attractes significant attention from academic researchers who consistently analyse the factors influencing VC decision-making. For example, the management team is identified as a key factor contributing to investment success (Gompers et al., 2020). However, Kessler et al. (2012) argue that founder characteristics only influence initial funding, while the founding process itself, including co-founder collaboration and managing expectations, has a greater impact on long-term viability. Additionally, Lerner and Nanda (2020) point out how the decision-maker's background and characteristics can influence investment choices. These factors hold varying weights in the selection process, potentially influenced by industry, company age, prior success, and geographical location (Gompers et al., 2020). Moreover, VCs' selection of investments can significantly influence the value creation of target startups by providing them with access to networks, mentorship, and crucial growth capital (Gompers et al., 2020).

The growing availability of data and advancements in computational power encourage the adoption of artificial intelligence (AI) and machine learning (ML) by VC firms. This trend enables VCs to leverage these technologies for enhanced data analysis and more informed investment decisions. This presents a compelling opportunity to explore the application of ML across various frontiers within the VC landscape, including optimisation of investment strategies and data-driven funding decisions (Cumming et al., 2022). Scholars are actively investigating the use of AI and ML to predict the future of private companies, focusing on bankruptcy risk and potential exit routes (Bhat and Zaelit, 2011; Arroyo et al., 2019). Krishna et al. (2016) leverage factors relevant to funding rounds to predict startup success and failure using supervised learning models like Support Vector Machines and Random Forests. Additionally, researchers use ML features to analyse the importance of investment candidate criteria. Corea et al. (2021b) introduce the Early-stage Startups Investment (ESI) framework, which utilises Gradient Tree Boosting to examine the influence of founders' demographics, professional backgrounds, psychological characteristics, and co-founder dynamics.

Beyond supervised learning models, researchers explore alternative approaches to address challenges in VC decision-making. Dellermann et al. (2017) propose a hybrid model that combines machine intelligence with human expertise to evaluate the qualitative aspects (soft signals) often present in startup success. Additionally, multi-criteria decision-making (MCDA) frameworks utilising fuzzy theory (Minola and Giorgino, 2008; Zhang, 2012; Afful-Dadzie et al., 2015) and goal programming (Aouni et al., 2013) employs to simulate the VC evaluation process and predict funding decisions. Unsupervised learning has limited application in understanding relationships and classifying startups into industry domains based on text descriptions (Kharchenko et al., 2023). Similarly, Xiong and Fan (2021) implement a semi-supervised method to analyse the VC network structure and identify industry leaders. Although supervised learning currently dominates VC research, this study advocates for a shift towards reinforcement learning (RL) as a promising alternative for portfolio recommendation.

5.2.2 Reinforcement learning for venture capital portfolio recommendation

Reinforcement learning (RL) presents a distinct machine learning paradigm compared to supervised learning. While supervised learning relies on labelled datasets, RL employs an agent that interacts with its environment through trial and error. This framework is successfully applied to simulate financial market participant interaction and optimise investment strategies for maximising long-term portfolio returns (Charpentier et al., 2023). Many practitioners and scholars investigate and propose the application of RL models in diverse financial domains, including stock trading (Azhikodan et al., 2019), stock index and commodity futures contracts (Deng et al., 2017), hedging instruments (Bühler et al., 2018), asset allocation (Moody and Saffell, 2001), and the simulation of limit order book markets for market makers (Spooner et al., 2018).

While reinforcement learning (RL) showcases its success in liquid markets like public equities, its application in the private capital market, characterised by illiquid assets with investment horizons of eight to ten years, remains underexplored. This limited and delayed feedback on portfolio performance hinders the agent's learning process and ability to optimise VC investment strategies. However, recent advancements in reinforcement learning-based recommender systems (RLRS) offer a promising avenue for addressing these challenges. RLRS can potentially be employed to simulate investment scenarios and identify investment opportunities that align with a VC's investment strategy and maximise long-term portfolio return. By leveraging recent research in RLRS, this study aims to evaluate the efficacy of RLRS as an alternative to traditional RL for portfolio selection in the private capital market.

Traditional recommender systems typically fall into two categories: collaborative filtering and content-based filtering. Collaborative filtering leverages relationships between products and user interests to generate recommendations based on user similarity or item characteristics. Conversely, content-based filtering utilises item descriptions to identify products that align with established user profiles. However, these traditional approaches exhibit limitations in capturing the nuances of sequential and dynamic user interactions. RL offers a compelling solution by incorporating continuous user engagement as feedback from the environment. This enables the formulation of recommendation problems as Markov Decision Processes (MDPs) (Afsar et al., 2022). The iterative nature of RL allows the agent to continuously refine its policy through interaction with the environment, potentially leading to superior recommendations that surpass those solely reliant on user ratings or static training data (Chen et al., 2019). A prominent application of RL in recommender systems lies in web recommendations. This approach empowers the system to dynamically interact with users, framing recommendations as actions rather than solely relying on static user behaviour gleaned from historical web usage data (Taghipour et al., 2007). In addition, recommender systems employing RL can recommend sequences of items, such as a playlist of songs (Liebman and Stone, 2014), or leverage multi-MDP tasks to capture user-specific attributes (Lei and Li, 2019).

This study aims to formulate an MDP framework to optimise startup recommendations within the VC investment domain. This framework, named VC-RLRS, is a novel model designed to evaluate top startup recommendations. Formulating this recommendation problem as an MDP requires consideration of several key design choices. These include state representation, state representation, which provides the agent with information about the current investment environment; exploration strategy, which determines how the agent balances between exploring new actions and exploiting existing knowledge; and reward function, which provides feedback to the agent, guiding its decision-making and learning process. Furthermore, crucial parameters such as the exploration rate, which governs the balance between exploring new ventures and exploiting known successful investments, and the discount factor, which determines the relative importance of short-term and long-term rewards, significantly influence the agent's decision-making process. To investigate the impact of these design choices, the study explores the following hypotheses:

> H_1: Including historical recommendations within the state representation will outperform those relying solely on current data in recommending high-growth potential startups, as measured by average return.

H_2: A reward function emphasising long-term financial returns will outperform other functions recommending high-growth potential startups, as measured by average return.

H_3: A higher exploration rate, which encourages the exploration of new startups alongside established investment strategies, will outperform a lower exploration rate that prioritises exploiting known successful ventures in recommending high-growth potential startups, as measured by average return.

H_4: A higher discount factor, which influences the agent's prioritisation of future rewards, will outperform a lower discount factor that emphasises short-term gains in recommending high-growth potential startups, as measured by average return.

By examining the influence of these design choices through experimentation, this research aims to contribute to the development of a novel VC-RLRS model for identifying high-growth potential startups.

5.2.3 Generalist vs. Specialist investment strategies in venture capital

Venture capital funds often focus on startups that align with their specific investment strategies. These strategies consider factors such as the stage of funding (e.g., pre-Series B), preferred industries, and geographical locations. By adhering to these criteria, venture capital funds can attract investors with similar goals and make informed deal selections. This approach raises the ongoing debate about the optimal level of portfolio concentration for balancing performance and risk (Norton and Tenenbaum, 1993).

Generalist VC funds, by definition, offer more diversification in startup selection, allowing them to adapt to a rapidly evolving market and avoid overexposure to a single sector. However, maintaining a knowledgeable board team and analysing deals in unfamiliar markets can be resource-intensive. Furthermore, generalist funds encounter competition from specialist VC funds with a strong track record. These specialist funds can be more alluring to entrepreneurs, as they often possess dedicated resources specifically tailored to support value creation within their chosen market niche. This targeted approach can potentially surpass the capabilities of a generalist fund in terms of value creation for startups in a particular industry (Gabbert et al., 2022).

However, specialist funds also face potential limitations. A downturn in the target market can significantly impact a fund's portfolio due to a lack of diversification within a single industry. Furthermore, focusing on a specific sector might lead to overlooking promising startups in other industries, potentially hindering the identification of high-return investment opportunities.

The ongoing debate about the relative performance of generalist and specialist venture capital funds remains inconclusive. While PitchBook's report (Gabbert et al., 2022) found no significant differences in technology and healthcare, research by Gompers et al. (2009) suggested that incorporating experienced specialists can enhance overall fund performance. This aligns with the arguments of Bygrave (1987, 1988) regarding the benefits of leveraging technical, product, and market expertise to mitigate risk.

The proposed VC-RLRS model can be employed to evaluate its effectiveness in selecting investments across specific domains of interest, such as technology, healthcare, and fintech. This investigation leads to the following hypothesis:

> H_5 Venture capital funds that limit their startup recommendations to a single industry will underperform funds that consider startups across multiple industries, as measured by average return.

It is important to acknowledge that various fund types exist, such as fund-of-funds, bridge funding, and mezzanine funding, which are established for specific company stages. However, these types fall outside the scope of the current research. The VC-RLRS framework has the potential to be adapted for formulating recommendations within both generalist and specialist VC funds. Furthermore, incorporating additional technical components, such as deep learning, can be explored to enhance the model's performance. A more detailed discussion of these elements is provided in the methodology section.

5.3 Methodology

This research proposes a novel application of Reinforcement Learning for Recommendation Systems, termed VC-RLRS, to simulate VC investment decision-making. The VC-RLRS model formulates the interaction between the investor and startup entities as a Markov Decision Process (MDP). By framing the selection process as an MDP, the agent can learn optimal investment strategies and recommend a list of the top ten startups[1] with high exit potential and exceptional returns. This approach has the potential to streamline the VC due diligence process by efficiently identifying promising investment opportunities.

5.3.1 Fundamentals of reinforcement learning

The recommendation problem can be formalised as an MDP characterised by a five-component tuple (S, A, P, R, γ)

1. State Representation (S): Represents the set of all possible states that define the agent's current situation within the environment.

2. Action (A): A set of actions that the agent can perform in each state. These actions influence the transition to the next state and the potential reward received.

3. Transition probability (P): This probability is denoted as $P(s'|s,a)$ and reflects the likelihood of transitioning from state s to state s' after taking action a.

4. Reward function (R): This function, $R(s,a)$, represents the immediate feedback the agent receives after taking action a in state s. The agent's goal is to learn a policy that maximises the total expected reward (or value) over time.

5. The discount factor (γ): This parameter $\gamma \in [0,1]$, determines how the agent values future rewards. A higher γ emphasises the importance of long-term outcomes, while a lower γ signifies a stronger focus on immediate rewards.

Reinforcement learning offers a diverse range of model architectures suited to different applications. Model-based RL allows the agent to learn an internal

[1] Average number of portfolio companies for the medium-sized fund.

representation of the environment, facilitating future planning capabilities. Policy optimisation, which refers to the process of determining the optimal probability of taking action a in state s, can be achieved through either tabular or function approximation methods. However, this approach is not well-suited for the VC investment domain, where the complete set of possible states is often unavailable for the agent to learn from. This limitation necessitates the use of model-free RL, such as Q-Learning, making it the more appropriate and practical choice for the proposed VC-RLRS model.

5.3.2 Q-Learning

Q-learning (Watkins and Dayan, 1992) is an RL algorithm particularly well-suited for the VC investment domain due to its model-free, value-based, and off-policy nature. Unlike model-based approaches that require a learned model of environment dynamics, Q-learning directly interacts with the environment to learn optimal actions. This is crucial in VC investment where the full environment dynamics are often intricate and not readily available. Furthermore, Q-learning's focus on the value function, which estimates the long-term reward an agent can expect from taking a specific action in a given state, aligns well with the objective of identifying high-growth potential startups. Through trial and error, the agent learns by experiencing the consequences of its actions, effectively bypassing the need for explicit transition models, which are often impractical in this context.

Within the VC-RLRS model, Q-learning employs an iterative approach across a predefined number of episodes. In each episode, the agent begins in a randomly chosen state, defined by the features of available startups. To navigate this state space, an exploration strategy guides the selection of the next state (next recommended startup). The policy leverages the Q-table to determine the optimal next action (selecting the next recommended startup). It achieves this by selecting the startup with the highest Q-value, reflecting its estimated long-term reward potential. Notably, Q-learning is an off-policy algorithm, meaning that the data used to update the Q-table can come from a different policy than the one used for exploring the environment. Once the optimal action for the next state (the highest Q-value) is chosen, the Q-table is updated using a specific temporal difference (TD) formula, incorporating the Bellman equation (Formula 5.1) to estimate the expected value based on the current state, previous state, and chosen action.

$$Q^{new}(s,a) \leftarrow Q(s,a) + \alpha * [R(s,a) + \gamma * max(Q'(s',a')) - Q(s,a)] \tag{5.1}$$

The Q-learning updates rule incorporates key parameters that influence the learning process specific to VC investment decisions. The core update equation utilises the concept of temporal difference, where the agent refines its understanding of the value associated with taking an action in a particular state. This is achieved by considering the immediate reward (denoted by $R(s,a)$) received from taking action a in state s, along with the estimated future reward obtainable from the next best state-action pair (represented by $Q'(s',a')$). $Q(s,a)$ denotes the current Q-value associated with the current state-action pair. Here, $max(Q'(s',a'))$ reflects the agent's estimate of the maximum expected future reward achievable from the next possible states and actions.

There are other key parameters that influence the learning process: learning rate (α) and discount factor (γ). The learning rate ($\alpha \in [0,1]$) determines the weight given to newly acquired information compared to existing Q-values. A learning rate of $\alpha = 0$ implies no updates based on new experiences, hindering learning. Conversely, $\alpha = 1$ completely replaces prior knowledge with new information, potentially leading to instability. The discount factor ($\gamma \in [0,1]$) reflects the agent's intertemporal preference, balancing the value of immediate rewards (e.g., potential for high investment return in the short term; $\gamma = 0$) with the importance of future rewards (e.g., long-term growth potential; $\gamma = 1$).

5.3.3 Deep Q-Learning

Reinforcement learning can leverage deep neural networks to effectively address complex decision-making tasks, including recommendation problems. Deep learning (DL) offers a significant advantage in uncovering non-linear relationships within high-dimensional state spaces. However, this enhanced capability comes with trade-offs, including high computational demands and potentially less interpretable results (Mousavi et al., 2018).

Several studies explore the potential of deep learning-based reinforcement learning (DRL) for financial applications. For instance, Hu and Lin (2019) explore the application of Deep Reinforcement Learning (DRL), which leverages deep neural networks to effectively manage the high dimensionality of state spaces encountered in financial tasks like stock portfolio management. DRL can keep track of historical states and actions while also searching for optimal parameters during policy optimisation. Deng et al. (2017) propose Deep Direct Reinforcement Learning (DDRL), where a deep learning component automatically captures market conditions and performs feature engineering before feeding data to the RL module. This approach demonstrates promising results in real-time stock and commodity futures trading. Empirical evidence from Park et al. (2020); Gao et al. (2020); Jiang et al. (2017) underscore the advantage of DRL for portfolio management. Their research showcases that deep learning-based RL models can outperform traditional trading strategies and be applied to multi-asset portfolios. Furthermore, Liu et al. (2022) develop a DRL library specifically designed for implementing stock trading strategies. This library simplifies development by providing pre-configured trading environments and constraints, facilitating backtesting of DRL-based strategies.

The majority of research on DRL applications in finance focuses on publicly listed equities, where asset prices naturally lend themselves to time-series analysis. Private capital, however, deals with illiquid assets, resulting in non-time series data for portfolio companies. This distinction presents a unique opportunity to explore the application of DRL to private capital management, specifically focusing on how deep learning can be leveraged to achieve optimal recommendation policies for VC investment decisions. The integration of DRL enables the model to extract and leverage nuanced features from the state representation, leading to more informed decision-making when recommending startups for both generalist and specialist VC funds.

This study leverages Deep Q-Network (DQN) (Mnih et al., 2013), an RL technique that incorporates a deep neural network architecture on Q-Learning. This Q-network is trained to approximate the Q-value function, which maps state-action pairs to the expected cumulative future rewards associated with taking a specific action within a given state. A significant advantage of DQN lies in its scalability compared to traditional Q-learning methods that rely on Q-tables. DQN effectively addresses the curse of dimensionality, a challenge encountered in RL when the size and complexity of Q-tables grow exponentially with increasing environmental complexity. This capability allows DQN to handle high-dimensional state representations, making it particularly well-suited for this research setting that involves intricate investment decision-making within a multi-dimensional feature space.

During this training, the agent evaluates startups and stores these experiences as tuples of $(state, action, reward, next\_state)$ within an experience replay buffer. This buffer serves as a repository of historical interactions, providing valuable data for the Q-network to learn from. The Q-network itself receives the current state (representing the available startups) as input and outputs Q-values for each possible action (selecting a startup). To optimise the learning process, the network leverages loss functions based on the temporal difference (TD) error (Equation 5.2). This concept, rooted in the Bellman equation (Equation 5.1), guides the network towards minimising the discrepancy between the predicted Q-values and the target Q-values. In essence, the network iteratively refines its decision-making strategies by adjusting its internal weights through gradient descent. During training, experiences are randomly sampled from the replay buffer to facilitate the continuous update of the Q-network. This implementation adheres to the Deep Q-Learning algorithm outlined in Appendix J.

$$Loss(\theta) = ((r + \gamma * max_{a'} Q'(s', a'; \theta^{-})) - Q(s, a; \theta))^2 \tag{5.2}$$

5.4 Experiment design

The design of Markov Decision Processes (MDPs) within RLRS applications necessitates careful consideration of several factors to ensure optimal performance in the specific problem domain. Afsar et al. (2022) explore various design considerations, including state representations, exploration strategies, and reward functions. These elements significantly impact the agent's learning process and decision-making capabilities.

In the context of VC investment, state representation and actions are crucial for defining the environment the agent interacts. The proposed VC-RLRS model represents the state and actions as a set of available startups that the agent can evaluate and recommend. Next, the exploration strategy plays a vital role in balancing exploitation (leveraging existing knowledge) and exploration (discovering new possibilities). This balance is essential for guiding the agent's learning process and identifying high-potential startups that may not be readily apparent. Finally, the reward function serves as a key signal for the agent, providing feedback on the desirability of its actions. In the VC-RLRS model, the reward function can be designed to consider various factors relevant to VC investment decisions, such as potential return on investment, exit potential, and alignment with the investment strategy. By carefully crafting these reward signals, the agent can be guided towards recommending startups with characteristics that align with the VC's investment goals.

To identify the optimal configuration for the VC-RLRS model, the study evaluates various combinations of these design elements within the context of the specific problem statements. This evaluation process allows for fine-tuning the agent's learning process and ultimately enables it to recommend a list of the top ten startups that maximise the expected cumulative return during the investment period.

5.4.1 Designs of state representation

Leveraging the existing knowledge of diverse state representation approaches in venture capital (VC) investment, this study proposes three distinct state design configurations. This exploration aims to identify the most effective configuration for capturing the critical information necessary for informed VC investment decisions.

- **State Design A**: This configuration represents the simplest state design, where the state, denoted by s_t, solely consists of the features of a single startup. The action, a_t, corresponds to the selection of that particular startup for the portfolio (Figure 5.1). This design mimics the scenario where an investor iteratively evaluates and adds startups to their portfolio, one at a time.

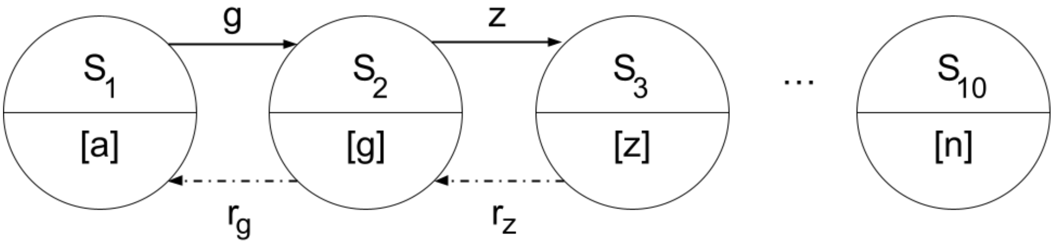

Figure 5.1: State design A visualisation: incorporating features of a single startup.

- **State Design B**: Building upon State Design A, this configuration incorporates a broader state representation, denoted by s_t. Inspired by the work of Taghipour et al. (2007); Liebman and Stone (2014), it includes the list of previously recommended startups alongside the features of the current startup under consideration, as shown in Figure 5.2. This expanded state representation introduces a memory component, allowing the agent to consider its recommendation history and potentially avoid suggesting the same startups repeatedly.

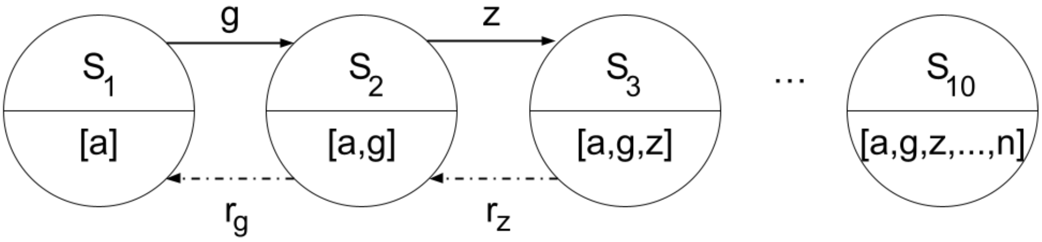

Figure 5.2: State design B visualisation: incorporating a list of recommended startups as memory.

In state S_1, the agent initially selects startup a randomly. It then recommends startup g and receives a reward r_g for this action. The agent transitions to the next state with the updated list of recommended startups, $[a, g]$. This process continues until the agent reaches state S_{10} with a list of ten recommended startups.

- **State Design C**: This configuration incorporates the most comprehensive state representation, denoted by s_t It includes the current investment year, and the list of all recommended startups to date as illustrated in Figure 5.3. This design reflects the assumption of an annual deal flow processing cycle, where the agent can select and recommend a single startup per year to be added to the portfolio.

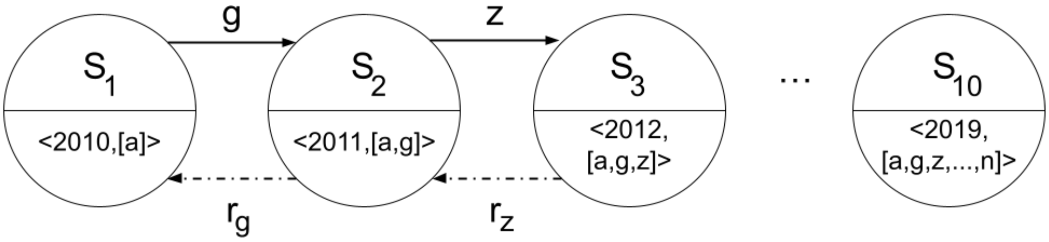

Figure 5.3: State design C visualisation: including investment year and recommended startups.

This design assumes a ten-year investment horizon, ranging from 2010 to 2019. In the final state, S_{10}, the agent has a list of ten recommended startups along with the corresponding investment years.

Within each episode, the initial state consists of a single, randomly chosen startup. The agent iteratively selects subsequent startups until it generates a list of the top ten recommendations. During this exploration process, the Q-table is updated to reflect the reward associated with each action taken within a given state. The size of the Q-table typically scales with the number of possible states and actions. In this case, the initial state and action space are identical, reflecting the number of startups the agent can recommend. However, this can vary depending on the chosen state representation design (as detailed previously).

For State Design A, the Q-table maintains a fixed size of $number\_startups \times number\_startups$. This is because both the number of states and actions correspond to the total number of available startups. However, State Designs B and C operate under the assumption that the order of startup selections does not influence the reward function. Consequently, the Q-table becomes dynamic and expands during each training episode. The number of states grows based on the cumulative list of recommended startups encountered by the agent. In this scenario, the action space reflects the next possible recommended startup, leading to a new state defined by the updated recommendation list

5.4.2 Exploration strategy

The exploration strategy is a critical component influencing the agent's state transitions within the VC investment search space. Q-learning offers various exploration strategies, each with specific parameters that guide the agent's action selection process. This process necessitates a balance between exploration (seeking new, potentially high-growth investment opportunities) and exploitation (leveraging the agent's accumulated knowledge to maximise expected returns). This study evaluates the performance of the following three exploration strategies (Sutton and Barto, 2018):

- **Epsilon-Greedy strategy**: This widely adopted exploration strategy employs an exploration rate (denoted by ϵ). With a probability of ϵ, the agent prioritises exploration by selecting a random action, fostering the discovery of potentially valuable yet underexplored investment opportunities within the search space. Conversely, with a probability of $1-\epsilon$, the agent leverages its current knowledge by selecting the startup associated with the highest Q-value. This action represents the most promising investment choice based on the agent's current understanding of the investment landscape.

- **Epsilon-Decay strategy**: Building upon the Epsilon-Greedy strategy, this approach incorporates a decaying exploration rate. This signifies that the agent explores more frequently in the initial stages, prioritising the discovery of novel investment opportunities. As the agent accumulates experience and refines its knowledge, the exploration rate gradually declines, favouring exploitation (selecting startups with the highest estimated returns based on the learned Q-values). This dynamic exploration approach allows the agent to balance the need for initial exploration with the goal of maximising long-term returns through informed investment decisions. The formula governing the decay of the epsilon value is presented in Equation 5.3.

$$\epsilon = \epsilon_0 * (decay\_rate^T) \tag{5.3}$$

- **Boltzmann (Softmax) strategy**: This strategy utilises the softmax function to compute action probabilities based on the Q-values of available startups. It incorporates a temperature parameter (denoted by τ) that modulates the level of exploration during the learning process. Higher temperature values correspond to increased exploration, promoting the selection of a wider range of startups and fostering the discovery of underexplored investment opportunities. Conversely, lower temperatures steer the agent towards exploitation by prioritising startups with the highest estimated returns, as reflected by their Q-values. The mathematical formula for calculating action probabilities using the softmax function is presented in Equation 5.4.

$$P_t(a) = \frac{exp(Q_t(a)/\tau)}{\sum_{i=1}^{n} exp(Q_t(i)/\tau} \tag{5.4}$$

5.4.3 Reward function

In the context of reinforcement learning, the reward function is crucial for guiding the agent's behaviour by providing feedback on its interactions with the environment. This study proposes and evaluates three distinct reward function designs, incorporating key VC success metrics such as exit opportunities (Bhat and Zaelit, 2011; Arroyo et al., 2019), return on investment (Korteweg and Sorensen, 2010), and portfolio diversification (Norton and Tenenbaum, 1993; Gompers et al., 2009). These reward functions aim to effectively guide the VC-RSRL agent in making informed and strategic startup recommendations.

- **Company Status**: This reward function emphasises the agent's ability to identify startups with favourable exit prospects, including mergers and acquisitions

(M&A) and initial public offerings (IPOs), while discouraging investment in failing companies. The design incorporates a static scoring system. Selecting a startup facing closure results in a significant penalty of -100. Conversely, successful exits are rewarded: $+100$ for M&A, reflecting a potentially faster return path for VCs due to streamlined regulatory processes compared to IPOs (as evidenced in Smith et al. (2011)). Startups that remain private but secure recent funding (within the past 3 years) also receive a positive reward of +50, acknowledging their potential for future growth and eventual exits.

- **Return**: Aligning with the core objective of VCs to maximise return on investment (ROI), this reward function incentivises the agent to prioritise the selection of startups with high gross return potential. The design adopts the gross return calculation method established by Korteweg and Sorensen (2010), which utilises the post-money valuation from the startup's first funding round and the pre-money valuation from its latest funding round as detailed in Equation 5.5.

$$gross\_return = \frac{latest\_pre\_money\_valuation}{first\_round\_post\_money\_valuation} \tag{5.5}$$

Due to potential limitations in the availability of comprehensive startup valuation data, this study incorporates the difference in raised funding amounts as a complementary reward factor (detailed in Equation 5.6). Both gross return and raised funding amount can be calculated on a round-by-round basis, reflecting the incremental funding stages of a startup. However, for simplicity and practical application within the investment timeframe of VCs who typically hold portfolio companies for 5-10 years, this study focuses on the disparity between the first funding round and the most recent one. This aligns with the VC investment model, where appreciating valuation over multiple funding rounds translates to a positive return on investment.

$$raised\_amount = \frac{(latest\_raised\_amount - first\_round\_raised\_amount)}{first\_round\_raised\_amount} \tag{5.6}$$

- **Startup similarity**:

  This reward function investigates the influence of portfolio company similarity on the agent's decision-making process. To assess the level of concentration and diversification within the recommended portfolio, the design calculates the business description similarity between the current state (recommended startup) and the next potential recommendation (represented by the next state). Given that startup descriptions are initially presented in text format, the study employs a text pre-processing step to convert them into a numerical representation suitable for machine learning algorithms. This study uses Sentence Transformer, a popular sentence embedding technique, specifically leveraging the pre-trained model *distilbert-base-nli-mean-tokens* (Reimers and Gurevych, 2019). This pre-trained model offers a fast and efficient method compared to other options, as it maps English sentences and paragraphs into a 768-dimensional dense vector space. Text similarity between two startups is then determined using the cosine distance metric (Equation 5.7) between their respective embedded vector representations.

$$startup\_similarity = cos(\theta) = \frac{A * B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.7}$$

5.4.4 Recommendation evaluation

Each episode simulates the complete investment decision-making process until a terminal state is reached, signifying the generation of a list of the top 10 recommended startups. To evaluate the performance of these recommendations, the study primarily employs the average raised funding amount (detailed in Equation 5.6). This metric aligns well with the practical consideration to evaluate the potential return of startups and it is readily accessible for a broader set of startups. The following is a breakdown of other relevant metrics for exit strategy and company status:

- **Acquisition rate**: This metric reflects the percentage of recommended startups that achieve successful exits through mergers and acquisitions (M&A). M&A exits are generally desirable due to their potential for high returns for the VC (Smith et al., 2011).

- **IPO rate**: This metric captures the proportion of recommended startups that experience successful exits via initial public offerings (IPOs). IPOs can generate significant returns for VCs upon public share offering.

- **Failure rate**: This metric represents the percentage of recommended startups that cease operations, resulting in a loss of investment capital for the VC.

- **Reinvestment rate**: This metric denotes the percentage of previously recommended startups that are selected again for reinvestment during the investment process. While a high reinvestment rate might indicate promising companies, excessive reinvestment can hinder portfolio diversification, a crucial principle for managing investment risk.

5.5 Data description

This study utilises startup data sourced from Crunchbase [2], a comprehensive database recognised for its extensive information on startups and private companies. Crunchbase offers valuable details such as funding history and investor profiles, facilitating informed investment decisions. However, a large volume of potential investment opportunities can pose a significant challenge for VCs in terms of efficient screening and due diligence. To mitigate this challenge, VCs often employ filtering mechanisms to identify opportunities aligning with their established investment strategy. Reflecting this practical approach, the current study focuses on UK-based startups founded between 2010 and 2020; this timeframe selection enables the exploration of a manageable dataset while maintaining relevance to contemporary investment landscapes.

The initial dataset consists of startup information obtained from Crunchbase. To address the scarcity of data with complete funding and performance details, data filtering is necessary. This process focuses on retaining UK-based companies with either company valuation or funding amount information. This filtering step resulted in a dataset of 3,382 companies. To establish a more manageable search space for the VC-RLRS agent, further data reduction is implemented through random sampling of 1,000 startups. Key descriptive statistics for these sampled startups, employed for subsequent evaluation purposes, are summarised in Table 5.1.

Table 5.1: The statistics summary of 1,000 startups used in the experiment of this study.

Name	Count	Mean	Std	Min	Max
Panel A: Startup Characteristics					
Age	1000	4.824	2.682	0.000	10.000
Located in London (Dummy)	1000	0.658	0.475	0.000	1.000
Operating status (Dummy)	1000	0.823	0.382	0.000	1.000
Closing status (Dummy)	1000	0.119	0.324	0.000	1.000
IPO exit status (Dummy)	1000	0.009	0.095	0.000	1.000

[2] Accessed on 5 March 2023 using the Crunchbase research access

M&A exit status (Dummy)	1000	0.049	0.216	0.000	1.000
Panel B: Funding					
Number of funding rounds	1000	2.231	1.455	1.000	12.000
Number of investors	1000	2.805	4.243	0.000	44.000
Total funding in USD (Log)	1000	13.944	1.916	6.908	21.107
Post-money valuation in USD (Log)	466	15.050	1.550	11.082	22.292
Non-equity (Dummy)	1000	0.093	0.291	0.000	1.000
Angel and Crowdfunding (Dummy)	1000	0.297	0.457	0.000	1.000
Pre- and Seed (Dummy)	1000	0.689	0.463	0.000	1.000
Early-stage VC (Dummy)	1000	0.159	0.366	0.000	1.000
Later-stage VC (Dummy)	1000	0.003	0.055	0.000	1.000
Private equity (Dummy)	1000	0.016	0.125	0.000	1.000
Debt financing (Dummy)	1000	0.056	0.230	0.000	1.000

The analysis reveals several interesting trends within the dataset of 1,000 UK-based startups. As illustrated in Figure 5.4, the majority of these startups were founded between 2014 and 2016, with an average age of approximately 4.8 years. London emerged as the dominant location for these startups (65.8%), potentially due to its concentrated ecosystem of funding resources and talent pool.

In terms of operational status, the data indicates that the majority (82.3%) of the startups are still operational. Only a small percentage of them exit through M&A (4.9%) or IPO (0.9%). The analysis of funding stages reveals that the average startup undergoes two funding rounds, with at least two investors participating on average. Pre-seed and seed funding rounds are the most common (68.9%), followed by angel investor and crowdfunding rounds (29.7%), and early-stage funding (up to Series C) at 15.9%.

Interestingly, a smaller proportion (9.3%) of startups secures non-equity grants. Notably, later-stage funding (3%), private equity involvement (1.6%), and debt financing (5.6%) are less prevalent. These companies with later-stage funding can potentially represent successful growth trajectories and potential future exits via

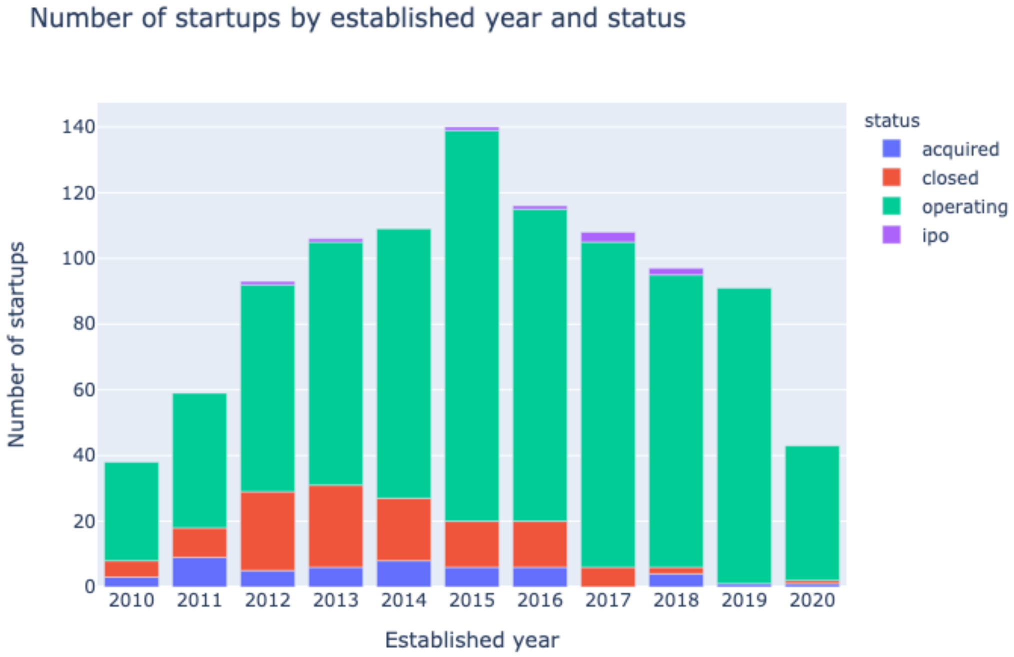

Figure 5.4: The number of startups established in each vintage year and grouped by the current exit and operating statuses.

M&A or IPO.

Finally, the data shows an average total funding of $11.9 million per startup, with an average post-money valuation of $38 million although valuation data is not available for all startups.

Specialised Investment by startups domains

This study emphasises the advantages of the specialist VC fund setting compared to generalist funds, as outlined in Section 5.2.3. Specifically, it examines the capability of the proposed VC-RLRS model to recommend startups within three key industries driving the global economy: Information Technology (IT), Financial Technology (FinTech), and Healthcare (Dealroom, 2024b). To assess the agent's performance and ensure alignment with real-world VC practices, the evaluation process utilises state representations on subsets of startups categorised by their industry domain. As previously mentioned, the startup data is segmented into the following industry groups for analysis:

- Information Technology (IT): 579 companies.

- Financial Technology (FinTech): 624 companies.

- Healthcare: 477 companies.

This industry-based breakdown facilitates a more comprehensive examination of the agent's decision-making capabilities within each specific sector.

5.6 Results

5.6.1 The effects of parameters

This section aims to examine the impact of key parameters within the VC-RLRS model on the agent's decision-making behaviour. Specifically, the study examines how exploration rates and discount factors influence the agent's ability to select actions that lead to high rewards, as outlined in hypotheses H_3 and H_4.

To isolate the effects of exploration rates and discount factors, the study fixed certain hyperparameters. The learning rate α is set to 0.01, controlling the weight assigned to new information during Q-value updates (Equation 5.1). Additionally, the decay rate employed in the epsilon-greedy exploration strategy (detailed in Equation 5.3) is configured to 0.01. This fixed decay rate ensures a gradual decrease in the exploration rate as the agent learns.

5.6.1.1 Exploration and Exploitation

A critical feature of RL models lies in their ability to manage the exploration-exploitation trade-off. In the context of VC investment, this translates to the agent's dynamic decision-making process. A higher exploration probability increases the likelihood of the agent selecting startups for evaluation without prior bias. This mirrors a VC's initial due diligence phase for early-stage companies, reflecting the need to discover promising investment opportunities within a vast landscape of potential ventures. This exploration phase is crucial for achieving long-term portfolio success, as it allows the agent to identify novel opportunities while maintaining a diversified portfolio. While a high exploration rate encourages the discovery of high-rewarding opportunities, excessive exploration can hinder the agent's ability to exploit established and successful investment choices.

This section analyses the influence of the exploration rate on the agent's decision-making process within the VC-RLRS model as specified in hypothesis H_3. The study evaluates the average funding raised by the startups recommended by the agent across 1,000 episodes, each consisting of ten recommendations. To investigate this effect, the study configured hyperparameters specific to each exploration strategy employed by the VC-RLRS model. The epsilon-based strategy employed epsilon (ϵ) values of 0.25, 0.50, and 0.75. In addition, the study utilised the softmax strategy with temperature (τ) of values 1 and 5. The temperature in the softmax strategy controls the degree of exploration. To ensure consistency across experiments, the discount rate (γ) is fixed at 0.5. Subsequent tables present detailed results for the various epsilon values, state representation designs, and reward functions employed within the experiments.

The analysis of Table 5.2 reveals that the optimal epsilon value for maximising capital raised depends on the chosen combination of state representation design and reward function within the epsilon-greedy strategy. Although a higher epsilon value promotes exploration, potentially leading to the identification of high-performing startups, it does not guarantee success in maximising the funding amount raised. State Design C consistently favours exploitation across all reward functions, with an optimal epsilon value of 0.25.

Table 5.2: The results of different exploration rates $\epsilon \in [0.25, 0.50, 0.75]$ of epsilon-greedy exploration strategy across state representation designs and reward functions.

		Epsilon Value		
Design	Reward Function	0.25	0.50	0.75
A	Company Status	11.5±50.6	**17.1±53.6**	15.7±34.2
	Return	43.8±31.7	31.5±74.9	**45.0±99.5**
	Startup Similarity	14.9±46.8	**17.7±50.2**	17.3±36.0
B	Company Status	15.5±89.2	15.5±41.5	**17.7±42.7**
	Return	13.8±48.7	**15.8±38.2**	15.5±34.4
	Startup Similarity	**17.9±63.2**	15.1±39.1	17.5±38.1
C	Company Status	**16.2±17.4**	11.0±27.6	12.4±34.7
	Return	**21.2±19.1**	15.5±22.9	15.4±39.4
	Startup Similarity	**15.5±80.4**	8.9±14.8	12.3±28.7

Conversely, State Designs A and B demonstrate a dependence on the chosen reward function. When using the “company status” and “returns” reward functions, both designs achieve optimal performance with epsilon values between 0.50 and 0.75, suggesting a preference for exploration to identify potentially high-performing startups. However, the “startup similarity” reward function prefers exploitation for both designs, with optimal epsilon values ranging from 0.25 to 0.50. This contrast highlights the intricate interplay between state representation design, reward function selection, and the exploration-exploitation balance in optimising the agent’s decision-making process.

Table 5.3 reveals significant variations in performance compared to the epsilon-greedy approach. All States Designs exhibits a preference for exploitation when paired with the “company status” reward function. The lowest epsilon value (0.25) results in the highest average funding raised by the recommended startups. This suggests that for the "company status” reward, focusing on established venture characteristics might be more effective than exploration in identifying promising investment opportunities.

Table 5.3: The results of different exploration rates $\epsilon \in [0.25, 0.50, 0.75]$ of epsilon-decay exploration strategy across state representation designs and reward functions.

		Epsilon Value		
Design	Reward Function	0.25	0.50	0.75
A	Company Status	**12.6±87.8**	5.2±22.8	12.1±48.4
	Return	7.1±33.1	**17.8±122.1**	16.7±122.6
	Startup Similarity	13.7±115.8	24.8±45.0	**35.1±20.6**
B	Company Status	**16.2±125.5**	9.5±50.1	14.7±99.8
	Return	10.0±32.5	9.7±44.1	**11.5±85.1**
	Startup Similarity	**14.1±94.4**	8.6±35.8	12.2±89.4
C	Company Status	**4.1±0.3**	3.0±1.3	0.5±3.6
	Return	4.3±1.0	502.6±108.9	**522.2±236.0**
	Startup Similarity	6.5±0.9	6.5±1.5	**6.6±2.0**

State Designs A and B showcase contrasting preferences for the exploration rate when using the “startup similarity” function. Interestingly, Design A favours a higher exploration rate, whereas Design B exhibits a preference for exploitation.

These contrasting findings underscore the intricate interplay between the chosen state representation design and the specific startup characteristics captured by each representation.

On the other hand, State Design C yielded an exceptionally high average capital raised of 522.2. However, this result is accompanied by a considerably higher standard deviation compared to the performance with an epsilon value of 0.50. This highlights a potential trade-off such that a higher exploration rate might lead to the discovery of a single, high-performing startup but also introduce greater variability in the overall portfolio performance. However, its performance with other reward functions is generally low. For instance, the average funding raised with the "company status" and "startup similarity" functions is 2.53 and 6.5, respectively. These results suggest that State Design C might be less effective in capturing the relevant information needed for identifying promising startups when using these reward functions within the epsilon-decay framework.

Table 5.4 presents the results obtained with the softmax exploration strategy. Interestingly, a temperature value (τ) of 1 generally leads to a higher average capital raised compared to a temperature of 5. This suggests that a lower temperature, promoting exploitation, might be more effective in maximising capital raised across various reward functions. However, an important exception emerges with State Design B, where a higher exploration yields superior performance when using the "company status" and "startup similarity" reward functions. Additionally, significant differences in capital raised are observed between temperature values for State Designs A and C when the "return" function is employed. These findings highlight the continued interplay between state representation design, reward function selection, and the exploration strategy in influencing the agent's decision-making process.

Table 5.4: The results of different temperature $\tau \in [1, 5]$ of softmax strategy across state representation designs and reward functions.

		Temperature Value	
Design	Reward Function	1	5
A	Company Status	**18.6±44.7**	17.0±36.8
	Return	**41.1±87.3**	18.5±46.3
	Startup Similarity	**18.8±48.1**	15.7±32.0
B	Company Status	16.5±34.4	**18.1±43.0**
	Return	**18.9±42.2**	16.8±36.6
	Startup Similarity	16.4±34.0	**17.0±35.0**
C	Company Status	**17.3±31.5**	12.1±31.6
	Return	**77.2±50.3**	12.8±25.0
	Startup Similarity	**11.5±20.0**	11.4±18.8

Contrary to the initial hypothesis H_3, a higher exploration rate does not necessarily translate to a greater focus on identifying high-growth potential startups. The findings reveal a more complex interplay between three key factors influencing the agent's performance: exploration strategy, state representation design, and reward function selection. Notably, simply increasing the exploration rate to maximise the likelihood of identifying winner startups is not always the optimal approach for achieving superior portfolio performance.

5.6.1.2 Discount Factor

In conjunction with examining the influence of epsilon values on the agent's exploration rate, the study also investigates the impact of discount factors (γ) to confirm hypothesis H_4. Discount factors determine how the agent weighs future rewards against immediate rewards, reflecting an investor's focus on long-term returns in the context of venture capital investment.

To analyse the effect of discount factors across various configurations, the exploration rate (ϵ) is fixed at 0.50 for the epsilon-based strategy and the temperature value (τ) at 1.0 for the softmax strategy. This isolation of the discount factor's effect allows for a clearer understanding of its influence on the agent's decision-making process within different state representation designs and reward functions.

Table 5.5 presents the findings on the influence of discount factors within the epsilon-greedy strategy. The results suggest that higher discount factors generally lead the agent to prioritise long-term reward maximisation. This aligns with the real-world investment strategy of venture capitalists who focus on achieving superior returns over extended time horizons.

Table 5.5: The results of different discount factors $\gamma \in [0.25, 0.50, 0.75]$ of epsilon-greedy exploration strategy across state representation designs and reward functions.

		Gamma Value		
Design	Reward Function	0.25	0.50	0.75
A	Company Status	14.0±32.1	17.1±53.6	**18.5±47.2**
	Return	30.5±52.0	31.5±74.9	**62.3±67.1**
	Startup Similarity	**20.1±60.6**	17.7±50.2	14.2±39.5
B	Company Status	15.8±45.0	15.5±41.5	**16.6±41.7**
	Return	17.8±48.9	15.8±38.2	**18.6±61.2**
	Startup Similarity	**16.8±53.0**	15.1±39.1	16.5±49.4
C	Company Status	10.0±21.5	11.0±27.6	**14.8±28.1**
	Return	12.2±24.7	15.5±22.9	**21.1±26.5**
	Startup Similarity	10.0±17.2	8.9±14.8	**14.6±25.5**

Interestingly, an exception emerges for State Designs A and B when using the "startup similarity" reward function. In this specific case, the lowest discount factor ($\gamma = 0.25$), which prioritises short-term rewards, yields the highest average capital raised. This deviates from the general trend observed and suggests that for these configurations, focusing on short-term gains might be more effective. The overall performance using the "startup similarity" function remains lower compared to the strategy using the "return" reward function across all state designs.

The impact of discount factors on the performance of the epsilon-decay strategy exhibits greater variability compared to the epsilon-greedy approach, as illustrated in

Table 5.6. The epsilon-decay strategy shows no consistent pattern across different discount factor values. One configuration, however, stands out: State Design C with the "return" reward function. A discount factor of 0.50 leads to a significantly higher average capital raised of 502.620 compared to other discount factors for this specific combination. The results suggest a high degree of variability in the influence of discount factors within the epsilon-decay framework.

Table 5.6: The results of different discount factors $\gamma \in [0.25, 0.50, 0.75]$ of epsilon-decay exploration strategy across state representation designs and reward functions.

		Gamma Value		
Design	Reward Function	0.25	0.50	0.75
A	Company Status	13.0±116.3	5.2±22.8	**18.3±142.4**
	Return	5.0±10.0	**17.8±122.1**	11.9±50.1
	Startup Similarity	9.2±32.0	**24.8±45.0**	9.7±40.4
B	Company Status	**11.4±88.6**	9.5±50.1	11.2±40.1
	Return	10.4±38.8	9.6±44.1	**18.6±124.4**
	Startup Similarity	**15.8±119.6**	8.6±35.8	15.2±111.4
C	Company Status	**4.7±1.3**	3.0±1.3	3.4±1.2
	Return	3.9±0.8	**502.6±108.9**	6.2±0.9
	Startup Similarity	6.5±1.0	6.5±1.5	**7.9±39.6**

Table 5.7 details the impact of discount factors within the softmax strategy. Interestingly, for State Design A, all reward functions achieve the highest average capital raised with a discount factor of 0.50. This suggests that within this specific configuration, a focus on medium-term rewards might be more advantageous compared to other discount factor values.

Table 5.7: The results of different discount factors $\gamma \in [0.25, 0.5, 0.75]$ of softmax exploration strategy across state representation designs and reward functions.

		Gamma Value		
Design	Reward Function	0.25	0.5	0.75

Table 5.7 continued from previous page

		Gamma Value		
A	Company Status	18.0±40.7	**18.6±44.7**	18.2±46.2
	Return	33.5±75.3	**41.1±87.3**	39.0±81.1
	Startup Similarity	18.1±42.6	**18.8±48.1**	18.7±46.4
B	Company Status	**17.2±36.4**	16.5±34.4	15.8±35.3
	Return	18.6±42.9	**18.9±42.2**	15.9±31.4
	Startup Similarity	17.5±37.7	16.4±34.0	**20.0±44.7**
C	Company Status	**24.9±19.1**	17.3±31.5	23.7±20.1
	Return	67.2±71.9	77.2±50.3	**89.2±79.9**
	Startup Similarity	11.2±18.2	**11.5±20.1**	11.1±25.8

State Designs B and C showcase a preference for short-term rewards when using the "company status" reward function. This is evidenced by the optimal discount factor of 0.25, suggesting that prioritising immediate returns is most effective. However, for these same state designs, the remaining reward functions perform best with discount factor values ranging from 0.50 to 0.75. This indicates that a focus on balancing medium-term and long-term returns becomes more prominent when the agent considers reward signals beyond just company status. These contrasting findings within State Designs B and C highlight the intricate interplay between state representation design, reward function selection, and discount factors in influencing the agent's decision-making process within the softmax strategy.

The analysis provides partial support for hypothesis H_4 such that higher discount factors generally influenced the agent to prioritise long-term reward maximisation. This aligns with the VC investment strategy by practitioners and potentially leads to strong performance in identifying top-tier startups.

5.6.2 The performance of design choices

Building upon our previous analysis of individual hyperparameters (exploration strategies and discount factors) on the VC-RLRS model's performance (Section 5.6.1), this section delves into the interplay between reward functions, exploration strategies, and state representations. By examining how these design choices interact, the study aims to gain a deeper understanding of their combined impact on the agent's decision-making process and the financial success of the constructed VC portfolios.

Reward Functions

To confirm hypothesis H_2, the analysis further investigates the reward functions detailed in Section 5.4.3 and categorises the results to identify the most effective configurations. Table 5.8 demonstrates that State Design C, when used with the epsilon-decay strategy and the "return" function, achieved the highest average capital raised of 522.2. This configuration is followed by those utilising the "startup similarity" and "company status" functions, respectively. State Design B, conversely, is consistently absent from the top performers across all reward functions, suggesting its relative ineffectiveness in identifying promising young companies.

Table 5.8: The results of startup recommendations performance grouped by the reward functions.

Reward Function	Design	Strategy	Mean±Std	Parameters
Company Status	C	Softmax	24.9±19.1	$\gamma = 0.25, \tau = 1$
Return	C	Decay	**522.2±236.0**	$\gamma = 0.50, \epsilon = 0.75$
Startup Similarity	A	Decay	35.1±20.6	$\gamma = 0.50, \epsilon = 0.75$

This observation suggests a potential synergy between the parameter optimisation criteria for prioritising percentage returns and the "return" reward function itself, lending support to hypothesis H_2. This alignment might explain the superior performance observed for this configuration compared to those using the discrete numerical values in the "company status" function and the inherent maximum value of one associated with the "startup similarity" function.

Figure 5.5 visually confirms the superior performance of the "return" function for state Designs A and C across 3,000 episodes. However, State Design B exhibits a less conclusive pattern, with the “startup similarity” function occasionally surpassing the “return” function.

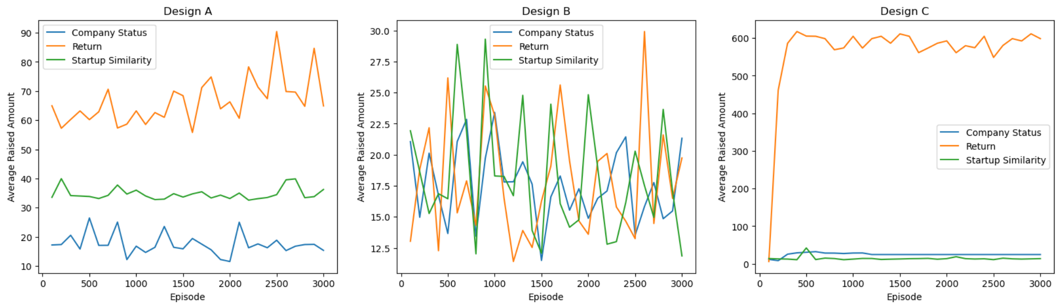

Figure 5.5: The line graph shows the performance of startup recommendations by the agent across 3,000 episodes by the reward functions and the state representation designs.

Exploration Strategy

When examining the relationship between exploration strategies and various reward functions (Table 5.9), a clear pattern emerges. The “return” reward function consistently outperforms the others in terms of average capital raised, regardless of the exploration strategy employed (epsilon-greedy, epsilon-decay, and softmax). This finding underscores the importance of aligning the reward function with the objective of maximising capital raised, as it significantly influences the agent’s decision-making process.

Table 5.9: The results of startup recommendations performance grouped by the exploration strategy.

Strategy	Design	Reward Function	Mean±Std	Parameters
Greedy	A	Return	62.3±67.1	$\gamma = 0.75$, $\epsilon = 0.50$
Decay	C	Return	**522.2±236.0**	$\gamma = 0.50$, $\epsilon = 0.75$
Softmax	C	Return	89.2±79.9	$\gamma = 0.75$, $\tau = 1$

The results in Table 5.9 suggest that grouping configurations by exploration strategy leads to superior performance compared to grouping by reward function (as seen in Table 5.8). This observation highlights the potential dominance of the exploration strategy in influencing the agent's decision-making process and ultimately impacting portfolio performance recommended by the VC-RLRS model.

Figure 5.6 provides a visual representation of how each exploration strategy performs within different state designs. State design C exhibits a clear preference for the epsilon-decay strategy, significantly outperforming all others within the first 500 episodes. The epsilon-decay strategy facilitates a two-phase decision-making process that aligns with VC investment strategies. During the initial exploration phase, the agent explores a broad range of startups. Subsequently, it gradually shifts focus towards exploiting previously identified ventures with high potential (exploitation phase). This result suggests a strong synergy between the decaying exploration rate of the epsilon-decay strategy and the information provided by State Design C. This alignment facilitates the agent to efficiently explore the search space and identify promising investment opportunities yearly.

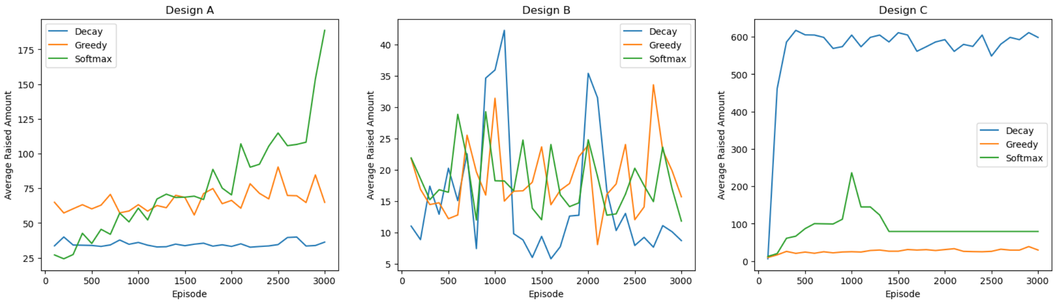

Figure 5.6: The line graph shows the performance of startup recommendations by the agent across 3,000 episodes by the exploration strategies and the state representation designs.

Conversely, State Design A demonstrates a continuous improvement in average capital raised when paired with the softmax strategy. This behaviour aligns with the core principle of softmax, where the exploration rate adapts based on the agent's experience. This allows for continuous exploration and potentially leads to better performance in the long run for state design A. In contrast, the epsilon-greedy strategy performs consistently but poorly with state design A. The fixed rate in epsilon-greedy might not allow for the level of exploration necessary to discover valuable investment opportunities within State Design A.

Consistent with prior observations from the reward function analysis (Table 5.5), state design B displays an inconsistent performance across exploration strategies (Figure 5.6). No single exploration strategy emerges as consistently superior for this design, as each occasionally outperforms the others at different points during the learning process.

State Representation

As outlined in hypothesis H_1, the choice of state representation is another critical factor that can influence the performance of the VC-RLRS model. The best-performing models across different exploration strategies, reward functions, and hyperparameters are summarised in Table 5.10 and Figure 5.7. State Design C exhibits superior performance, significantly outperforming all other designs within the first 500 episodes. This observation underscores the critical role of state representation in influencing the VC-RLRS model's decision-making process, leading to the maximisation of capital raised for young companies.

Table 5.10: The results of startup recommendations performance grouped the state representation designs.

Design	Strategy	Reward Function	Mean±Std	Parameters
A	Greedy	Return	62.3±67.1	$\gamma = 0.75$, $\epsilon = 0.50$
B	Softmax	Startup Similarity	12.0±44.7	$\gamma = 0.75$, $\tau = 1$
C	Decay	Return	**522.2±236.0**	$\gamma = 0.50$, $\epsilon = 0.75$

Interestingly, State Design B, which incorporates previously recommended startups into its state representation, yields lower performance compared to the memory-less State Design A. A potential explanation for this observation is that the specific information captured within State Design B's state might be less suitable for guiding the VC-RLRS's decision-making process. This finding warrants further investigation; exploring a range of evaluation metrics beyond capital raised, such as acquisition rates and portfolio diversification, can offer a more comprehensive understanding of State Design B's capabilities.

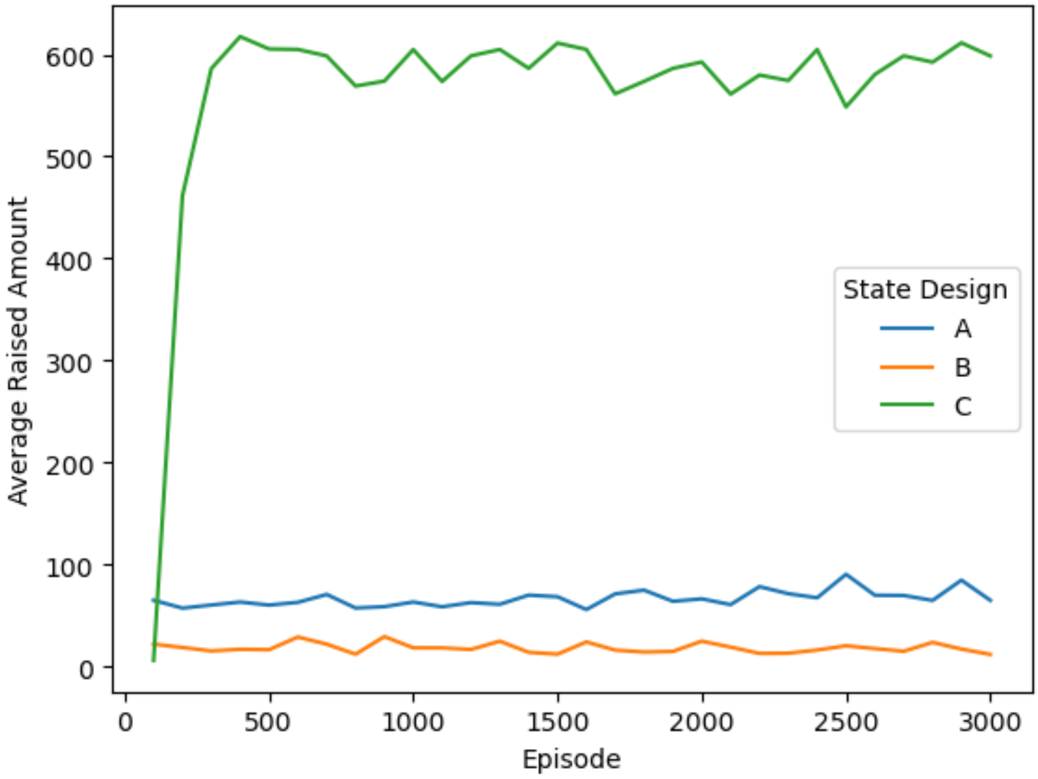

Figure 5.7: The line graph shows the performance of startup recommendations by the agent across 3,000 episodes by the state representation designs.

5.6.3 Comparative evaluation of proposed VC-RLRS

While average capital raised offers a valuable initial assessment, a more comprehensive evaluation of portfolio performance necessitates additional metrics. Examining previously identified relevant metrics, such as acquisition rate, IPO rate, failure rate, and reinvestment rate, provides a more comprehensive assessment of portfolio success. This broader analysis reveals whether State Design C, incorporating investment year and past recommendations, remains the top performer across various dimensions critical to VC portfolio success generated by the proposed VC-RLRS model.

The following visualisations are in the form of scatter plots, one for each state design. Each data point represents an episode of the top ten startups recommended by the VC-RLRS model. The data points are classified according to the target indicator, as indicated in the legend for each figure. The X-axis shows the average gross return of the recommended startups across those episodes. The Y-axis represents the average capital raised by the recommended startups within the same episode range. This visualisation allows us to examine potential relationships between state design, target indicators, and relevant metrics.

Acquisition rate

While exhibiting lower overall portfolio performance, State Design B displays a unique strength in identifying startups with the potential for exceptionally high returns through mergers and acquisitions (M&A) (Figure 5.8). State design B consistently identifies startups achieving the highest gross returns, even if the average capital raised for these startups is lower. A significant portion of the top-ten recommendations from state design B (43%) involve startups that exit through M&A. This observation suggests a potential benefit of incorporating information on previously recommended startups (memory) into the state representation. Specifically, this memory component might be particularly well-suited for uncovering promising, yet high-risk, M&A targets that traditional investment strategies might overlook.

In contrast to its strength in capital raised, state design C exhibits a decline in performance when considering M&A success rates. Notably, only a single episode within this design yields a successful M&A exit with significant returns. Additionally,

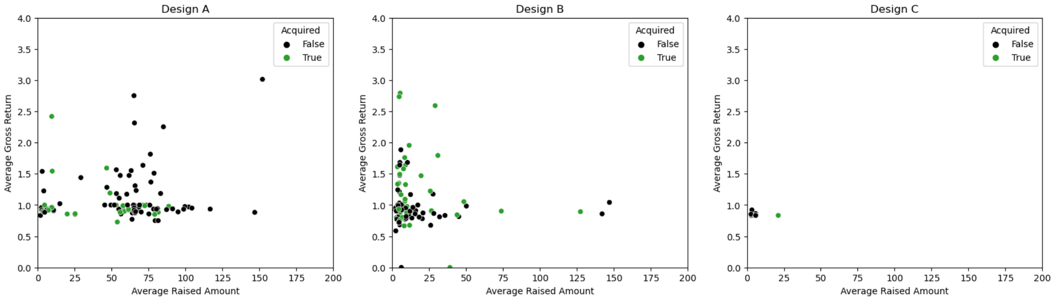

Figure 5.8: The scatter graph illustrates the relationship between average gross return and the raised amount of startups, as recommended by various state representation designs within the first 100 episodes. The data points are categorised by the mergers and acquisitions (M&A) status of the startups.

the overall results for state design C tend to cluster around a low average capital raised and gross return. Meanwhile, State Design A without the memory capacity achieves a success rate of 23% for episodes with at least one M&A exit. Additionally, State Design A maintained a level of capital raised comparable to other designs, suggesting it can balance the pursuit of high-growth ventures with the need for sufficient capital investment. This observation underscores the critical importance of a multifaceted approach to portfolio evaluation, moving beyond a single metric such as capital raised.

IPO rate

An analysis of IPO rates across recommended startups reveals a clear distinction between state designs as shown in Figure 5.9. Although the dataset exhibits a low overall IPO propensity, with only 0.9% of startups achieving IPO exits, State Designs A and B successfully identify high-growth ventures, evidenced by their selection of startups that subsequently went public at rates of 7% and 3%, respectively. Conversely, state design C does not recommend any startups that achieve IPOs within the observed timeframe.

Failure rate

Shifting the focus to failure rates, a critical metric for VC portfolio success, Figure 5.10 reveals a concerning trend for State Design C. The data suggests that a majority of the startups recommended by this design experience failure. This high failure rate

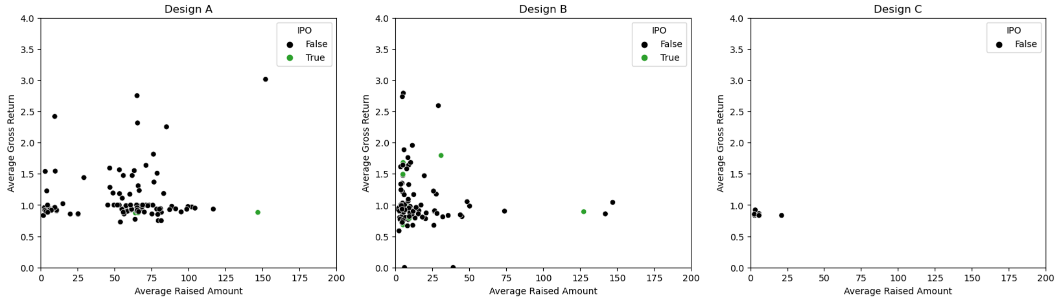

Figure 5.9: The scatter graph illustrates the relationship between average gross return and the raised amount of startups, as recommended by various state representation designs within the first 100 episodes. The data points are categorised by the IPO status of the startups.

highlights potential limitations in the agent's ability to learn from past experiences and avoid previously encountered unsuccessful ventures.

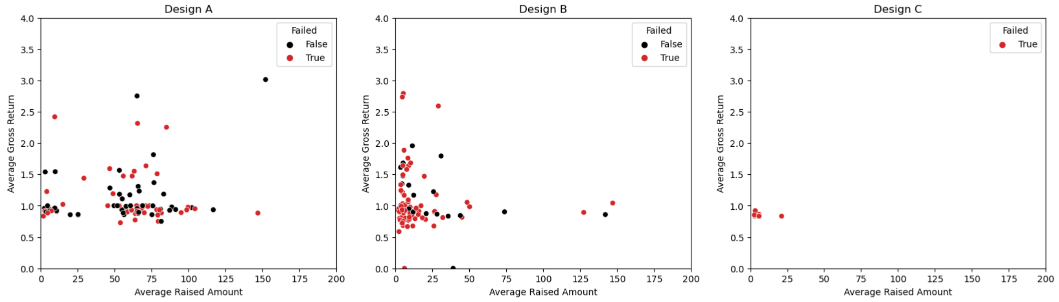

Figure 5.10: The scatter graph illustrates the relationship between average gross return and the raised amount of startups, as recommended by various state representation designs within the first 100 episodes. The data points are categorised by the failure status of the startups.

In contrast to State Design C, both State Designs A and B exhibit lower failure rates, with a smaller proportion of recommendations resulting in failed startups. However, the overall failure rate remains high across all designs, with at least one failure experience in 49% and 78% of episodes for Designs A and B, respectively. Despite exhibiting some degree of filtering for potentially resilient ventures, all state designs experience a concerningly high overall failure rate. This finding underscores the critical need for further research to improve the VC-RLRS model's capacity to mitigate portfolio failure and the consequent loss of investment.

Reinvestment rate

This section examines reinvestment rates, a key metric used in VC investment strategies to assess portfolio diversification. Diversification refers to the practice of spreading investments across a variety of companies or industries to mitigate risk. A high reinvestment rate, where the agent frequently selects the same ventures, can increase portfolio concentration and heighten exposure to a single company or industry.

Figure 5.11 reveals a distinct pattern in reinvestment rates across state designs. State Design B, which incorporates previously recommended startups into its state representation, demonstrates a significantly lower tendency to select the same ventures. Only 5% of recommendations in this design include at least one duplicated startup within the top ten list. This finding suggests that incorporating memory, as in design B, helps the VC-RLRS model avoid over-investing in the same companies, thereby promoting portfolio diversification and potentially reducing risk.

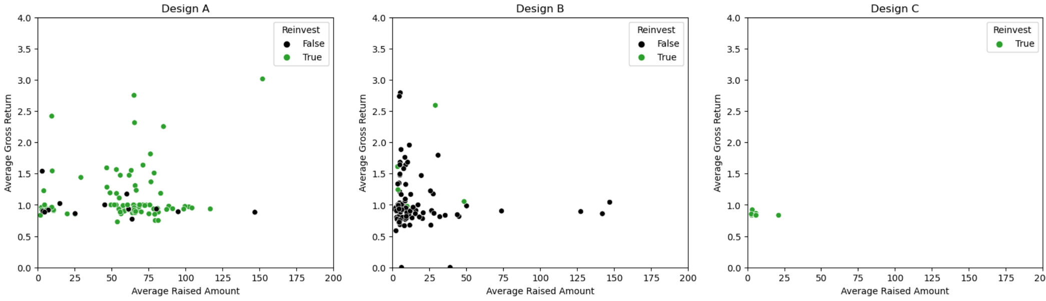

Figure 5.11: The scatter graph illustrates the relationship between average gross return and the raised amount of startups, as recommended by various state representation designs within the first 100 episodes. The data points are categorised by the reinvestment indictor

State design A, which does not incorporate previously recommended startups into its state representation, achieves a reinvestment rate of 13%. In contrast, State Design C, which incorporates both the investment year and a history of previously recommended startups into its state representation, exhibits a higher propensity for reinvesting in existing portfolio companies. All data points necessarily reflect this behaviour, indicating reinvestment within a ten-year investment timeframe. This behaviour aligns with some VC practices of pursuing follow-on investment rounds in promising ventures. However, this approach presents a potential trade-off between exploiting existing investment opportunities and mitigating concentration risk. While reinvestment can capitalise on prior successes, it can also lead to a less diversified portfolio and increased exposure to a smaller number of successful companies.

Example of VC-RLRS Recommendation

The following analysis delves deeper into the model's decision-making capabilities by examining top-performing episodes within each state design. Specifically, the analysis focuses on episodes that achieve both the highest average funding amount and the lowest failure rate. These episodes represent instances where the VC-RLRS model successfully balanced the pursuit of high-growth ventures with risk mitigation by identifying startups with both high potential returns and low failure probabilities.

Table 5.11 presents a sample recommendation from State Design A, which achieves the highest average capital raised among all designs (517.3). While this finding highlights the model's potential for identifying high-investment ventures, none of the startups experience an initial public offering (IPO) or merger and acquisition (M&A) exit. Second, the sectoral distribution of the recommendations within this sample leans towards the HealthTech and IT domains. Finally, the founding years of the recommended startups are concentrated after 2018. This observation suggests a potential preference for more recent ventures, warranting further investigation to understand the model's selection criteria across founding stages.

Table 5.11: The table presents the top 10 startup recommendations generated by an agent employing state representation design A. The agent utilises a greedy exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.75 and an exploration rate (ϵ) of 0.50.

<table>
<tr><th colspan="5">Episode: 1796 | Average Raised Amount: 517.3, Failure Rate: 0%, Reinvest Rate: 0%</th></tr>
<tr><th>Startup</th><th>Established</th><th>Gross Return</th><th>Raised Amount</th><th>Industry</th></tr>
<tr><td>C01</td><td>2018</td><td>0.8</td><td>N/A</td><td>Entertainment</td></tr>
<tr><td>U02</td><td>2018</td><td>N/A</td><td>2502.3</td><td>HealthTech</td></tr>
<tr><td>R03</td><td>2014</td><td>0.9</td><td>N/A</td><td>HealthTech</td></tr>
<tr><td>I04</td><td>2011</td><td>1.0</td><td>N/A</td><td>IT</td></tr>
<tr><td>D05</td><td>2019</td><td>N/A</td><td>300.8</td><td>FinTech</td></tr>
<tr><td>S06</td><td>2018</td><td>N/A</td><td>16.4</td><td>HealthTech</td></tr>
<tr><td>J07</td><td>2016</td><td>N/A</td><td>149.3</td><td>IT</td></tr>
<tr><td>F08</td><td>2019</td><td>N/A</td><td>25.8</td><td>EdTech</td></tr>
<tr><td>V09</td><td>2016</td><td>109.0</td><td>N/A</td><td>Cyber Security</td></tr>
</table>

Table 5.11 continued from previous page

Episode: 1796 | Average Raised Amount: 517.265, Failure Rate: 0%, Reinvest Rate: 0%				
W10	2012	0.3	N/A	Social Media

Despite achieving the lowest average capital raised among all state designs, State Design B offers a fascinating example of prioritising risk mitigation as shown in Table 5.12. This focus is evident in a sample recommendation showcasing a zero failure rate, even with lower average investment amounts 376.0 compared to other designs.

Table 5.12: The table presents the top 10 startup recommendations generated by an agent employing state representation design B. The agent utilises a softmax strategy with a reward function based on startup similarity. Key parameters for this strategy include a discount factor (γ) of 0.75 and a temperature (τ) of 1.

Episode: 1206 | Average Raised Amount: 376.0, Failure Rate: 0%, Reinvest Rate: 0%				
Startup	Established	Gross Return	Raised Amount	Industry
U02	2018	N/A	2502.3	HealthTech
M11	2013	0.8	N/A	News
A12	2018	0.9	N/A	FinTech
A13	2015	N/A	65.5	AI & Big data
O14	2018	N/A	58.9	FoodTech
E15	2013	N/A	0.5	HealthTech
O16	2015	N/A	3.9	HealthTech
I17	2012	N/A	0.2	FinTech
G18	2015	N/A	0.3	GreenTech
P19	2018	0.9	N/A	FinTech

The example recommended by State Design B focuses on the HealthTech and FinTech sectors. Interestingly, the model successfully identifies “U02” a startup within the healthcare sector that secures a significant funding round of 2,502.3, which is also selected by the state Design A. This finding suggests the model’s capability for identifying promising ventures within specific industries.

Furthermore, the inclusion of startup similarity within the reward function appears to yield positive results in terms of portfolio diversification. This is evidenced by the recommendation of “M11” from the entirely different news and publishing domain. This observation suggests that the model can balance sectoral focus with diversification, potentially mitigating risk through a broader portfolio composition.

Following the analysis of average capital raised, this section focuses on State Design C, which incorporates the investment year into the selection process. As shown in Table 5.13, sample episodes within this design achieves a funding raised amount of 630.0, demonstrating the highest performance among all state designs. However, 40% of the recommended startups are no longer operational. This finding underscores the potential shortcomings of relying solely on past returns as a predictor of future venture viability.

Table 5.13: The table presents the top 10 startup recommendations generated by an agent employing state representation design C. The agent utilises a decay exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.50 and an exploration rate (ϵ) of 0.75. A startup name that ends with an asterisk (*) means that the startup is no longer operating.

Episode: 127 | Average Raised Amount: 630.0, Failure Rate: 40%, Reinvest Rate: 20%					
Year	Startup	Established	Gross Return	Raised Amount	Industry
2010	K20*	2010	1.0	N/A	Blockchain
2011	U21	2010	0.5	N/A	FinTech
2012	I04	2011	1.0	N/A	IT
2013	I04	2011	1.0	N/A	IT
2014	R22*	2014	N/A	5.8	Social Media
2015	O23	2015	N/A	0.005	IT
2016	S24*	2016	0.1	N/A	Digital Media
2017	B25*	2016	0.8	N/A	Nanotechnology
2018	H26	2012	N/A	12.0	Hospitality
2019	U02	2018	N/A	2502.3	HealthTech

Furthermore, the presence of a duplicate recommendation (i.e., startup “I04” is selected in both 2012 and 2013) raises concerns regarding potential limitations in portfolio diversification within State Design C. While this duplication can be interpreted as a deliberate reinvestment strategy, it deviates from VC practices that typically avoid consecutive year investments in the same company.

Startup Coverage Analysis

To understand how agents navigate the extensive search space of 1,000 startups, this study analyses VC-RLRS's selection patterns. The 29 most frequently chosen startups by each agent are identified and visually illustrated in Figure 5.12.

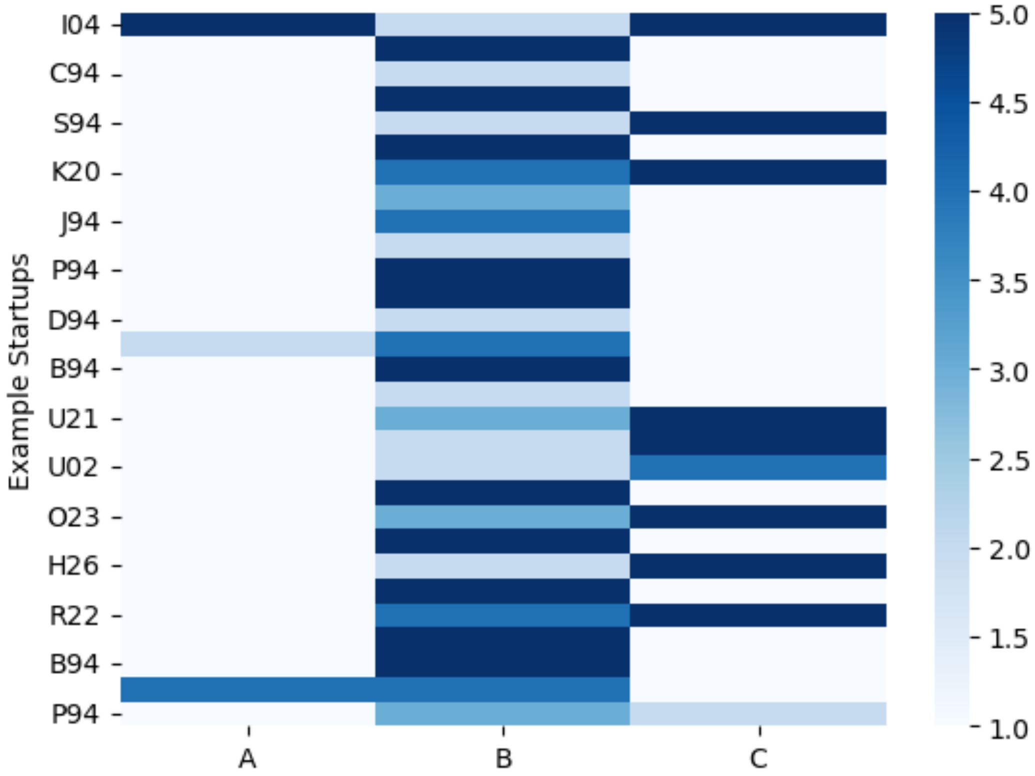

Figure 5.12: The heatmaps illustrate the selection patterns of the agents across the available startups. The colour intensity within each cell represents the frequency at which a specific startup is recommended by the agent across 500 episodes.

An analysis of agent selection patterns reveals a potential bias within State Design A. This is evidenced by the strong colour intensity for startup "I04", indicating a significantly higher selection frequency compared to other ventures. Conversely, the weaker colour intensity for a broader range of startups suggests a lower selection probability. Overall, state Design A achieves a high exploration rate, leaving only

5.8% of the startups unexplored.

Similar to State Design A, State Design C exhibits a preference for a limited subset of startups within the search space. This is reflected by the higher concentration of dark blue colours in Figure 5.12, indicating a significantly higher selection frequency for a specific group of ventures. Furthermore, the finding that 97% of the available startups remain unselected by this state representation underscores a potential concern regarding limited exploration within the search space. The selection patterns suggest that State Design C exhibits a risk of convergence to a local maximum within the search space. Prioritising a single startup might lead the agent to overlook the exploration of alternative ventures that can offer promising returns.

In contrast, State Design B emphasises a more balanced selection pattern within the search space. This is evidenced by the distribution of colour intensity across the figure, which suggests a more even exploration of the available startup options. Additionally, the low percentage (0.3%) of unselected startups by this model reinforces this observation. This finding suggests State Design B's effectiveness in navigating the search space and identifying a broader set of potential investment opportunities compared to the other designs analysed.

These findings underscore the inherent trade-off between exploration and exploitation in this context. While venturing into new areas of the search space (exploration) is crucial for identifying potentially high-performing startups, it also increases the risk of encountering unsuitable ventures. Conversely, focusing on a limited number of known options (exploitation) offers a safer route but may lead to missed opportunities.

A comprehensive evaluation of portfolio success requires a multifaceted approach that extends beyond the sole consideration of average capital raised. For example, State Design C, despite its impressive average capital raised of 522.2, demonstrates the importance of other factors in assessing overall portfolio performance. It struggles to identify promising ventures and exhibits concerning failure and reinvestment rates, hindering diversification. In contrast, State Designs A and B display a more balanced performance across metrics. As shown in Figure 5.13, they achieve a balance between high average gross return and funding amounts, while avoiding concentration risk and investment losses seen in State Design C. In response to hypothesis H_1, State Design B emerges as the strongest candidate for balancing all criteria due to its incorporation of the history of recommended startups. Nevertheless, further refinement is necessary to enhance its performance.

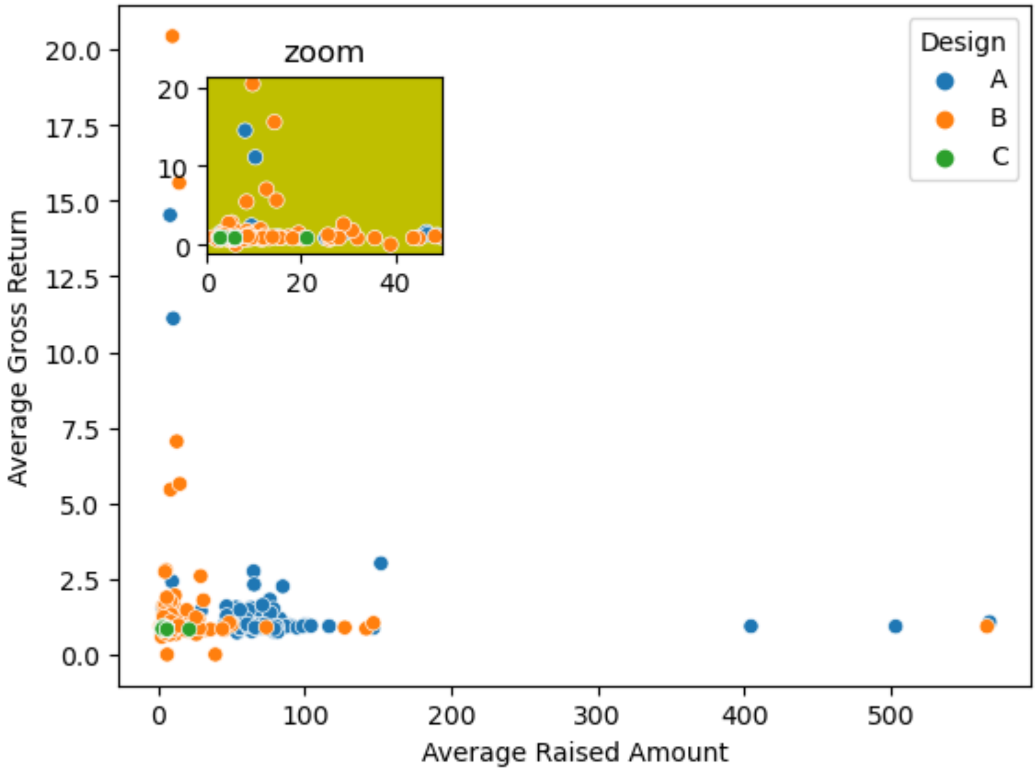

Figure 5.13: The scatter graph illustrates the relationship between average gross return and the raised amount of startups, as recommended by various state representation designs within the first 100 episodes. The data points are categorised by the state representations

5.6.4 Startup domains

To test hypothesis H_5, this section examines the agent's performance on datasets representing specific startup domains, as outlined in Section 5.5. The analysis focuses on three key sectors: Financial Technology (FinTech), Information Technology (IT), and Healthcare.

The results presented in Table 5.14 reveal that State Design A, the simplest form of our state representation design, achieves the highest overall performance. The healthcare sector exhibits the strongest average capital raised (145.6), followed by FinTech (78.3) and IT (63.0). Figure 5.14 visually reinforces these findings, illustrating State Design A's initial lead across all domains over 3,000 episodes. These findings can be attributed to the possibility that the dataset is filtered by specific industries, allowing the agent to prioritise the selection of successful ventures without the need for a memory function to maintain diversification.

Table 5.14: The table presents the performance of startup recommendations for various industry domains categorised by state representation design. The reward function, exploration strategy, and parameters employed here are identical to those detailed in Table 5.10.

	State Design		
Domain	A	B	C
FinTech	**78.3±183.0**	28.0±103.1	3.1±0.6
Healthcare	**145.6±164.4**	22.4±49.6	56.2±37.5
IT	**63.0±90.8**	25.2±58.0	14.8±1.2

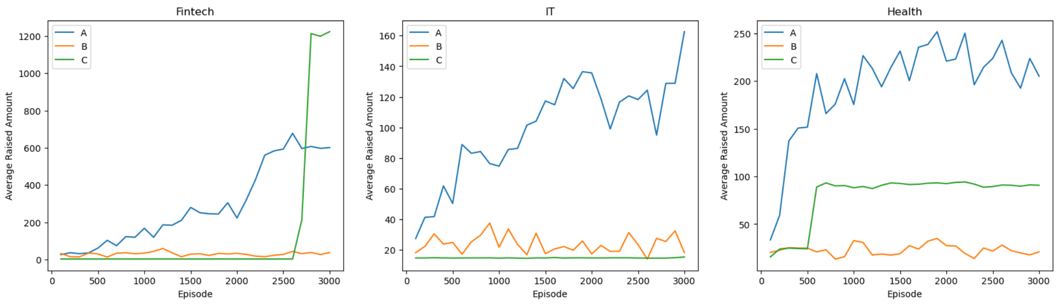

Figure 5.14: The line graphs show the performance of startup recommendation by the agent over 3,000 episodes, categorised by the state representations.

While State Design A achieves the highest average capital raised initially, State Design C eventually surpasses Design A in the FinTech sector after approximately 2,500 episodes. This observation suggests that the epsilon-decay strategy employed by State Design C, which encourages exploration before exploitation, can outperform a purely greedy strategy like State Design A in the long run, particularly within specific industry contexts. The FinTech domain, characterised by its dynamic nature and potentially higher risk-reward profiles, might be more receptive to this exploration-oriented approach. This finding highlights the potential benefits of balancing exploration and exploitation for long-term success in certain investment landscapes.

Tables 5.15, 5.16, and 5.17 showcase sample episodes within each domain (FinTech, IT, and Healthcare) that achieve both the highest average capital raised and the lowest failure rate. These results demonstrate the agent's ability to identify promising startups with strong financial performance and low risk of failure within specific industry sectors.

Table 5.15: The table presents the top 10 FinTech startup recommendations generated by an agent employing state representation design A. The agent utilises a greedy exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.75 and an exploration rate (ϵ) of 0.50.

Episode: 2517 | Average Raised Amount: 1266.3, Failure Rate: 0%, Reinvest Rate: 0%				
Startup	Established	Gross Return	Raised Amount	Subdomain
O27	2019	0.8	N/A	Compliance
C28	2016	0.5	14.0	Payments
B29	2014	105.6	N/A	Blockchain
T30	2010	N/A	6198.2	SMEs Finance
Q31	2018	1.7	N/A	Investment
B32	2015	0.9	N/A	Art
S33	2011	N/A	13.4	Cloud Computing
T34	2016	N/A	98.8	Payments
F35	2017	N/A	7.3	Investment

Table 5.15 continued from previous page

Episode: 2517 | Average Raised Amount: 1266.3, Failure Rate: 0%, Reinvest Rate: 0%				
L36	2013	2.4	N/A	E-commerce

Table 5.16: The table presents the top 10 IT startup recommendations generated by an agent employing state representation design A. The agent utilises a greedy exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.75 and an exploration rate (ϵ) of 0.50.

Episode: 1574 | Average Raised Amount: 509.9, Failure Rate: 0%, Reinvest Rate: 0%				
Startup	Established	Gross Return	Raised Amount	Subdomain
E37	2018	0.8	N/A	Augmented Reality
C38	2018	N/A	0.2	Software
D39	2019	0.9	N/A	AI & Big Data
H40	2011	0.9	N/A	Healthtech
T41	2018	0.6	N/A	Fashion
B42	2016	N/A	1.1	Construction
U02	2018	N/A	2502.3	Healthtech
R43	2020	N/A	4.5	Software
H44	2017	N/A	41.3	Cyber Security
R45	2019	0.9	N/A	AI & Big Data

Table 5.17: The table presents the top 10 Healthcare startup recommendations generated by an agent employing state representation design A. The agent utilises a greedy exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.75 and an exploration rate (ϵ) of 0.50.

Episode: 2351 | Average Raised Amount: 515.6, Failure Rate: 0%, Reinvest Rate: 0%				
Startup	Established	Gross Return	Raised Amount	Subdomain
C46	2014	0.7	N/A	E-commerce
H47	2018	N/A	522.0	Biotechnology
A48	2016	N/A	2.8	Diagnostics

Table 5.17 continued from previous page

Episode: 2351 | Average Raised Amount: 515.6, Failure Rate: 0%, Reinvest Rate: 0%				
E49	2015	N/A	1.0	Biopharma
H50	2014	1.9	N/A	Nutrition
D51	2016	0.9	N/A	Nutrition
U02	2018	N/A	2502.3	Fitness
F52	2012	0.9	N/A	Nutrition
B53	2010	N/A	0.3	Insurance
P54	2019	N/A	65.2	Pharmaceutical

Similar to limitations observed earlier, the VC-RLRS model prioritises maximising average funding but struggles to identify startups with high exit potential through M&A and IPOs, potentially missing out on valuable investment opportunities. Furthermore, measuring diversification within the context of individual industries presents a challenge. A breakdown analysis, potentially involving the segmentation of industries into subdomains, can provide a more precise assessment of portfolio diversification.

Figure 5.15 illustrates the selection frequency across all startups, revealing a concerning trend. During the initial 500 episodes, the agent exhibits a preference for a limited subset of startups within each domain, as evidenced by the strong colour gradient. This observation suggests a potential for limited exploration, particularly in the early stages of the learning process.

The observed behaviour of limited exploration during the initial learning phase can be attributed to the chosen hyperparameter settings, particularly the discount factor (γ) of 0.75 and exploration rate (ϵ) of 0.50. These settings may lead the model to prioritise exploitation by focusing on previously encountered startups rather than exploring new ventures. Within the first 500 episodes, the agent fails to select a significant portion of startups across all domains (57.5% in FinTech, 59.6% in IT, and a concerningly high 68.6% in Healthcare). These findings highlight the critical role of careful hyperparameter tuning in achieving a balance between exploration

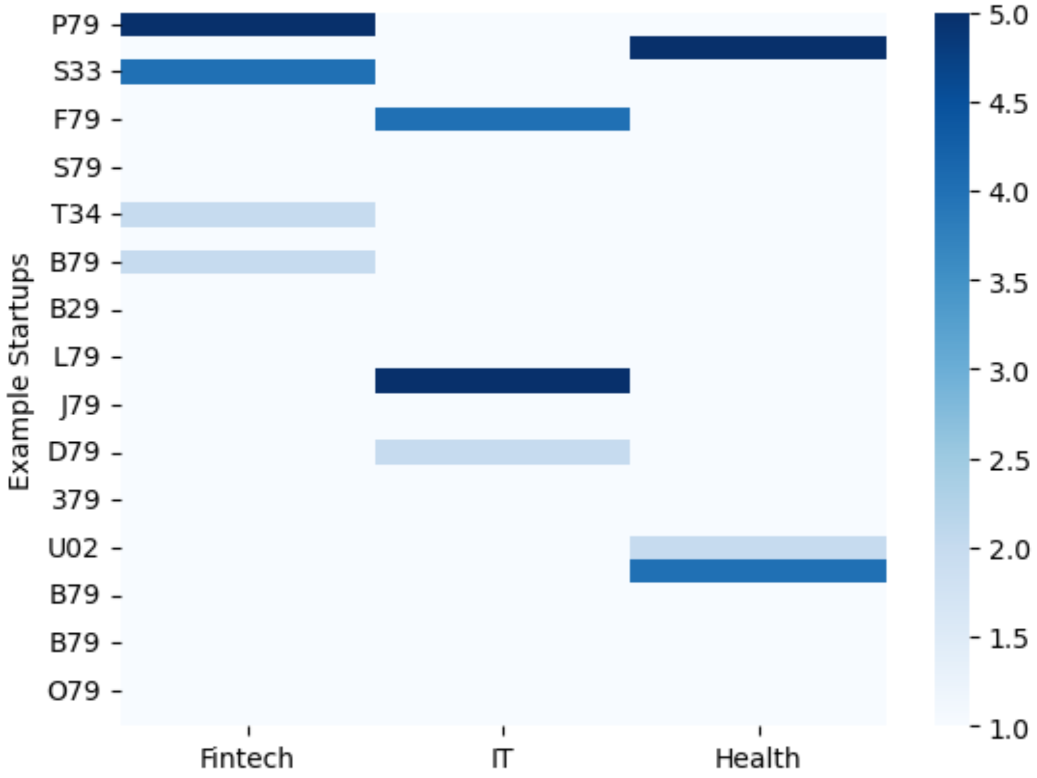

Figure 5.15: The heatmaps illustrate the selection patterns of the agents integrating state presentation A across the available startups. The colour intensity within each cell represents the frequency at which a specific startup is recommended by the agent across 500 episodes.

and exploitation, particularly when dealing with smaller datasets. Therefore, the findings support the hypothesis H_5 such that specialising the investment into a specific industry may hinder the portfolio performance compared with the generalised fund.

5.6.5 Deep Q-Network

This study proposes a novel hybrid approach that combines deep learning architectures with reinforcement learning techniques to enhance the performance of startup recommendations. The Deep Q-Network (DQN) model leverages the epsilon-decay exploration strategy, identified as the most effective across all evaluated models (Table 5.9). This strategy prioritises the exploration of the vast search space in the initial learning phase, gradually shifting towards the exploitation of known successful ventures as the model learns. To maintain consistency across evaluations, various reward functions and hyperparameter configurations are tested for the DQN model within each of the three state representation designs, while the epsilon-decay exploration strategy remains constant.

Despite incorporating a deep learning architecture, the DQN model does not achieve significant performance improvements. As shown in Table 5.18 and Figure 5.16, the mean average capital raised across all state design configurations fell below that of the baseline VC-RLRS model presented in Section 5.6.2. Notably, the DQN with state design A achieves the highest average capital raised amount (24.0), followed by designs C (11.0) and B (8.5).

Table 5.18: The table presents the performance of startup recom- emendations categorised by state representation design. with Deep Q Network.

Model	Strategy	Reward Function	Mean±Std	Parameters
A-DQN	Decay	Startup Similarity	24.0±28.2	$\gamma = 0.50, \epsilon = 0.75$
B-DQN	Decay	Return	8.5±33.8	$\gamma = 0.75, \epsilon = 0.50$
C-DQN	Decay	Return	11.0±21.6	$\gamma = 0.50, \epsilon = 0.75$

While prior analyses suggest State Design C's superiority, the DQN employing State Design A (A-DQN) exhibits a surprising initial outperformance. This is evidenced by the steeper upward trend in the A-DQN's performance curve during the initial episodes, as shown in Figure 5.17.

While the A-DQN model initially exhibits a promising upward trend in average capital raised, this early success is overshadowed by a concerning decline in

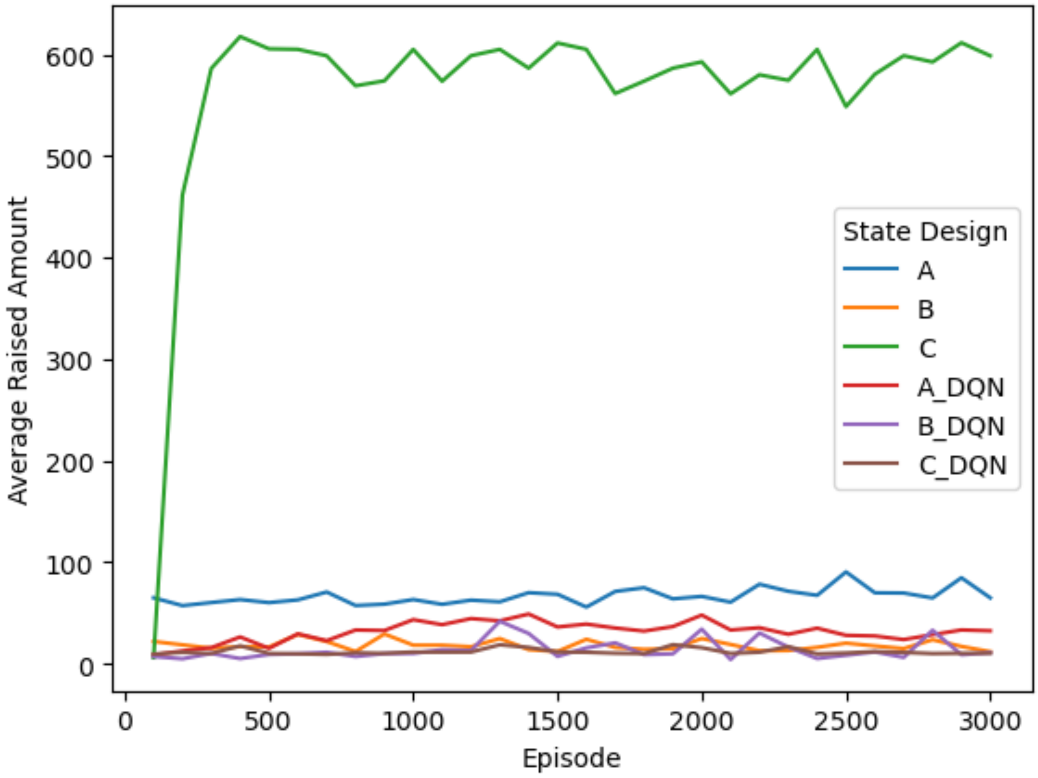

Figure 5.16: The line graph shows the performance of startup recommendations by the agent over 3,000 episodes, categorised by the state representation designs.

performance after approximately 1,500 episodes. This pattern suggests potential limitations within the A-DQN architecture, particularly when coupled with State Design A, for maintaining long-term performance. One possibility is that the agent may be overfitting to exploit readily available ventures, hindering its ability to generalise and explore new startups, potentially high-performing opportunities, as the learning process progresses.

The DQN model employing state design B (B-DQN) exhibits a distinct performance pattern compared to A- and C-DQN. The B-DQN model demonstrates a steady rise in the average capital raised, culminating in a peak around episode 1,000. However, this initial promise is followed by a period of performance volatility throughout the remaining episodes. Meanwhile, the C-DQN achieves a more consistent average amount raised in the range of 10 to 20 across all episodes.

A further analysis is conducted to examine the initial recommendations generated by each DQN-integrated state representation design. The analysis focuses on episodes

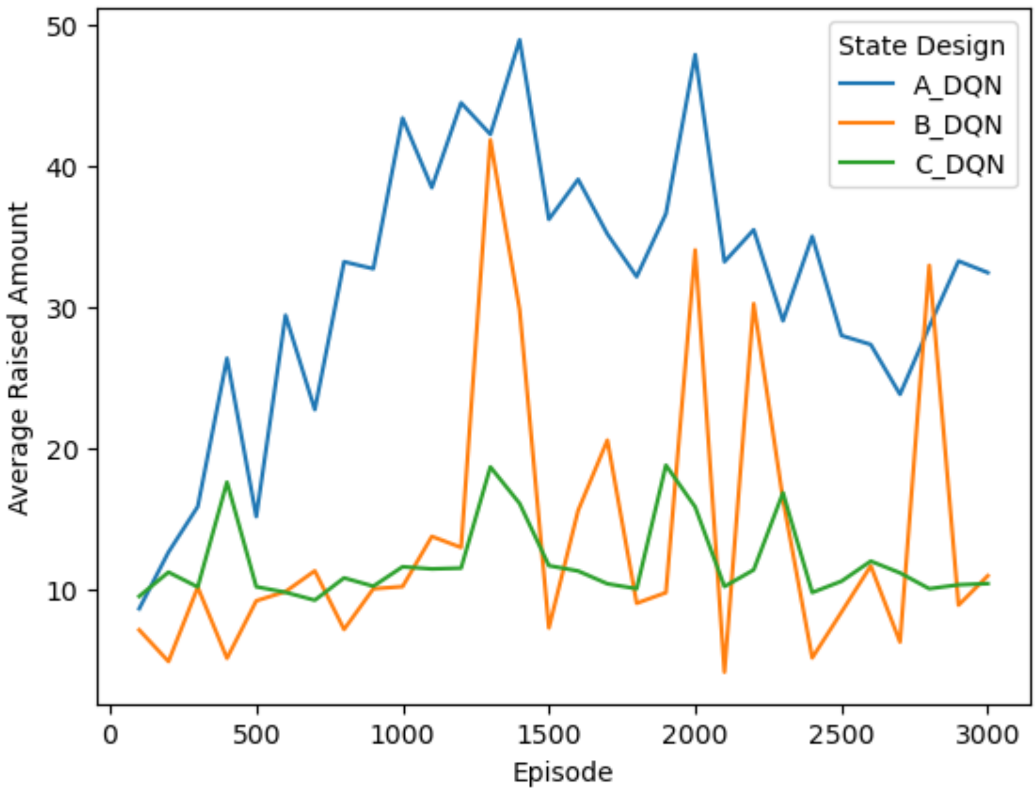

Figure 5.17: The line graph shows the performance of startup recommendations by the agent over 3,000 episodes, categorised by the state representation designs with the Deep Q Network.

achieving the highest average capital raised and minimal failure rates. However, the results presented in Tables 5.19 and 5.20 of State Designs A and B yield significantly lower average capital raised compared to the non-DQN models, at 13.4 and 4.0, respectively. Additionally, the failure rate remains high, particularly for State Design B at 40%. Unlike the baseline VC-RLRS model, which successfully avoids selecting failed ventures by not incorporating historical startup information into its decision-making process, the DQN-based models appear to struggle in this aspect despite leveraging recommendation history within the state representation. Furthermore, a significant reinvestment rate in both designs suggests a bias towards existing ventures, potentially hindering diversification.

Table 5.19: The table presents the top 10 startup recommendations generated by an agent employing state representation design A with Deep Q Networks. The agent utilises a decay exploration strategy with a reward function based on startup similarity. Key parameters for this strategy include a discount factor (γ) of 0.50 and an exploration rate (ϵ) of 0.75. A startup name that ends with an asterisk (*) means that the startup is no longer operating.

Episode: 1 | Average Raised Amount: 13.4, Failure Rate: 10%, Reinvest Rate: 20%				
Startup	Founded Year	Gross Return	Raised Amount	Industry
T55	2012	0.7	N/A	Edtech
T56*	2016	0.9	N/A	Social Network
S57	2015	N/A	16.7	E-commerce
K58	2018	N/A	4.2	Fintech
P59	2018	N/A	24.9	Fintech
L60	2014	N/A	10.1	Legal
K61	2015	2.1	N/A	Foodtech
T56*	2016	0.9	N/A	Social Network
T62	2017	N/A	14.7	Advertisement
H63	2015	N/A	9.6	Human Resources

Table 5.20: The table presents the top 10 startup recommendations generated by an agent employing state representation design B with Deep Q Networks. The agent utilises a decay exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.75 and an exploration rate (ϵ) of 0.50. A startup name that ends with an asterisk (*) means that the startup is no longer operating.

Episode: 2 | Average Raised Amount: 4.0, Failure Rate: 40%, Reinvest Rate: 30%				
Startup	Established	Gross Return	Raised Amount	Industry
F64	2013	1.0	N/A	Social Media
H65	2019	0.9	N/A	Education
H66*	2014	0.8	N/A	Fintech
I04	2011	1.0	N/A	IT
E67*	2013	0.6	N/A	AI & Big data
I04	2011	1.0	N/A	IT

Table 5.20 continued from previous page

Episode: 2 | Average Raised Amount: 4.0, Failure Rate: 40%, Reinvest Rate: 30%				
R22*	2014	N/A	5.8	Social Media
I04	2011	1.0	N/A	IT
E68	2020	1.6	2.2	Environmental Consulting
B69*	2015	0.9	N/A	Fintech

In contrast to the limitations observed in state designs A- and B-DQN, the C-DQN model (Table 5.21) employing state design C achieved a positive outcome: a zero failure rate and a zero reinvestment rate. However, this is accompanied by a concerningly low average capital raised of only 8.0.

Table 5.21: The table presents the top 10 startup recommendations generated by an agent employing state representation design C with Deep Q Networks. The agent utilises a decay exploration strategy with a reward function based on return. Key parameters for this strategy include a discount factor (γ) of 0.50 and an exploration rate (ϵ) of 0.75. A startup name that ends with an asterisk (*) means that the startup is no longer operating.

Episode: 3 | Average Raised Amount: 8.0, Failure Rate: 0%, Reinvest Rate: 0%

Year	Startup	Established	Gross Return	Raised Amount	Industry
2010	O70	2019	N/A	9.9	Quantum Computing
2011	D71	2010	1.2	N/A	AI & Big data
2012	L72	2011	N/A	0.4	E-commerce
2013	N73	2012	N/A	0.6	Cloud computing
2014	B74	2010	0.9	N/A	Publishing
2015	M75	2014	N/A	7.7	Healthtehch
2016	O16	2015	N/A	3.9	Healthtech
2017	P76	2016	0.9	N/A	Human Resources
2018	V77	2014	N/A	25.5	E-commerce
2019	W78	2016	0.9	N/A	AI & Big data

Although integrating deep neural networks (DNNs) into a reinforcement learning (RL) framework holds potential for startup recommendation, the observed performance may be limited by several factors. The dataset of 1,000 startups, while valuable, might not be sufficiently large for the DQN to effectively capture the complex relationships between the state representation and recommended startups, given the DNN's strength in handling extensive datasets with intricate features. Additionally, the choice of state representation, which provides the DQN with the necessary information, significantly impacts its performance. The current representations may not be optimally aligned with the DQN architecture, potentially limiting its ability to learn effective selection strategies. Despite integrating a deep learning architecture (DQN), this hybrid model does not exhibit a significant performance improvement compared to the VC-RLRS baseline.

5.7 Discussion

Venture capitalists face significant challenges in assessing and identifying high-potential investment opportunities to construct optimal portfolios that deliver exceptional returns. This complexity is often exacerbated by the illiquidity of private capital and the skewed distribution of returns, where only a few investments within a portfolio achieve high success (Cochrane, 2005). Such conditions, which intensify the already present information asymmetry (Denis, 2004), demand an evolution in investor decision-making criteria (Lerner and Nanda, 2020). While several studies explore machine learning, particularly supervised learning, for predicting company valuation and future performance, the application of reinforcement learning (RL), which uniquely enables agents to learn through trial and error, remains notably underexplored in this domain.

This research directly addresses this critical gap by successfully bridging reinforcement learning with venture capital, proposing a novel Venture Capital Reinforcement Learning Recommender System (VC-RLRS) designed for identifying high-growth potential startups. By leveraging recent advancements in RLRS, this study evaluates its efficacy as a promising alternative to traditional methods for portfolio selection within the private capital market. A key contribution lies in the development of a Q-learning model specifically tailored to the unique challenges of the venture capital industry, including its illiquidity and the limited data availability that reinforces information asymmetries. This innovative application extends the theoretical understanding of how adaptive learning systems can navigate highly uncertain and data-scarce financial environments (Afsar et al., 2022; Chen et al., 2019).

The state representation encoding the current conditions of the environment that agents interact with proves to be a crucial factor. Incorporating historical recommendations as a memory component, inspired by the work of Taghipour et al. (2007); Liebman and Stone (2014) to simulate how the VCs consider the past investments while reducing reinvestment in the same companies, yields promising results. The proposed design shows potential in capturing ventures with potential exit opportunities through mergers and acquisitions (M&A) and initial public offering (IPO). This aligns with the established goals of VC investment, which aim to invest in young companies with high growth potential and the possibility of significant returns through exits. However, imposing an annual investment limit constrains the model's ability to identify startups with high exit potential and avoid unsuccessful investments. Moreover, restricting investments to one per year is unrealistic in the dynamic venture capital industry. Such a limitation can lead to missed opportunities in promising startups or force investments during market downturns, particularly given the heavily skewed distribution of returns towards highly successful ventures (Cochrane, 2005). Further improvements in state representation can be achieved by integrating additional data, such as VC preferences or startup characteristics, similar to Lei and Li (2019)'s work, to capture user-specific attributes during the recommendation process. This aligns with the growing emphasis on personalisation and domain expertise in RL applications.

Reward functions provide essential feedback guiding an agent's interactions within an environment. This study designs and tests three unique reward functions considering investment return, potential venture exits, and portfolio diversification. Results indicate that the reward function prioritising the return can overshadow crucial factors such as venture stability and market fit. To address these limitations, a balanced approach considering both short-term and long-term outcomes is necessary. Developing multi-objective reward functions that incorporate a broader range of success metrics is crucial for optimising investment strategies. Moreover, the study examines the impact of exploration rate and discount factor on the RL agent's learning process. While high exploration rates, intended to mimic VC screening practices by increasing the likelihood of discovering promising startups, do not consistently yield high returns, the long-term nature of private capital investments suggests that a higher discount factor can be advantageous. This aligns with the well-established exploration-exploitation trade-off in reinforcement learning, emphasising the need to balance the discovery of new opportunities with the exploitation of known successes. The selection of investments is crucial, as it also influences the long-term value creation of a portfolio (Gompers et al., 2020).

In addition to its original contribution of designing and implementing the VC-RLRS to showcase the capabilities of reinforcement learning in the venture capital domain, this research successfully demonstrates the model's ability to recommend target ventures within specific industries, including Financial Technology, Information Technology, and Healthcare. The study highlights the model's adaptability to both generalised and specialised investment strategies. To further enhance its applicability, future research should explore tailoring the policy and reward function to specific diversified dimensions, such as geolocation and investment stage. This would help identify the optimal level of portfolio concentration for balancing performance and risk (Norton and Tenenbaum, 1993) while also accounting for venture capitalists' domain expertise and network effects.

Finally, this research lays the groundwork for employing deep Q-networks (DQNs) within a reinforcement learning framework for VC startup recommendations. While DQNs offer potential advantages, scalability challenges arising from large datasets necessitate exploring alternative architectures. The current study's reliance on a 1,000-UK startup dataset highlights the need for models capable of handling larger and more diverse datasets, incorporating factors such as geolocation, startup stage, and industry, which vary across VC funds. Future research can prioritise developing scalable RL architectures capable of handling complex state representations and substantial data volumes. By addressing these limitations, practitioners can enhance their investment decision-making through the VC-RLRS framework.

5.8 Conclusion

This research addresses the critical challenges of venture capital investment, particularly in the screening and selection of promising startups. To address these challenges, the study proposes a novel VC-RLRS model tailored for this domain, utilising the Q-learning algorithm, which is underexplored in private capital research. This study fills this gap by comprehensively exploring various RL design choices, as outlined in the hypotheses, including state representation, exploration strategies, reward functions, and hyperparameters such as exploration rates and discount factors. These parameters significantly influence the agent's behaviour in searching for and selecting startups that align with specific investment goals.

The proposed model demonstrates promising results across both generalised and specialised investment strategies. It successfully recommends startups with high average funding amounts, suggesting their potential to attract future investment rounds. The study emphasises the importance of crafting an effective state representation that captures relevant information for the RL agent's decision-making. Additionally, it underscores the need for a tailored reward function design that balances short-term gains with long-term considerations crucial for VC success. The exploration-exploitation trade-off is also addressed, suggesting that a high discount factor, reflecting the long-term investment horizon of VC, might be beneficial.

Beyond its traditional supervised learning approach, which relies on labelled training data to predict future portfolio stages, the VC-RLRS model offers a promising tool with practical applications in venture capital. The study highlights ways to further develop the framework by addressing limitations like incorporating VC-specific data and refining the reward function. By leveraging an improved VC-RLRS, practitioners can make more informed investment decisions and potentially improve overall portfolio performance.

Chapter 6

Conclusion

This research delves into the inherent challenges of entrepreneurial finance, particularly at its early funding stages, with a high level of risk. This domain is uniquely complex due to the significant information asymmetry (Denis, 2004) prevalent in illiquid investments, compounded by limited historical data and less stringent regulatory requirements compared to public equity markets. Furthermore, the private capital lifecycle involves distinct stages, from initial screening and due diligence to active portfolio management aims at achieving successful exits and generating extreme returns. This illiquid market is characterised by a skewed distribution of returns, where significant success is typically driven by only a small number of high-performing assets within a portfolio (Cochrane, 2005). In this dynamic environment, investors must also be agile in responding to emerging trends, such as the innovation of blockchain technology and its token financing mechanisms, alongside the growing demand of sustainable finance that incorporates social and environmental factors into investment decisions. The three empirical chapters present in this thesis collectively address these fundamental challenges by leveraging the potential of alternative data and machine learning to support decision-making and enhance integration across the entire private capital investment lifecycle. The subsequent sections outline the implications of key findings, contributions, and future work derived from each chapter.

Chapter 3 highlights the emergence of blockchain technology and the subsequent demand for Initial Coin Offering (ICO) as a new venture financing method. Despite their growing popularity, the chapter identifies persistent challenges, particularly a lack of regulatory framework and existence of information asymmetry, which impede accurate token valuation. Building upon gaps in existing literature, particularly the limitations in token rating methodologies and the limited analysis of whitepaper content (Ofir and Sadeh, 2020; Florysiak and Schandlbauer, 2019; Ante et al., 2018), this study conducts a detailed examination of multiple signalling factors influencing post-ICO token returns. The key findings reveal a significant discrepancy between token ratings and realised six-month returns, with a misclassification rate of approximately 67%, which fundamentally challenges the reliability of token ratings. These ratings may indicate fundraising success (Bourveau et al., 2022), but not long-term ICO returns. In response to these critical limitations, this chapter investigates alternative data sources, such as whitepapers and social media, specifically from Twitter, constructing a custom-built ICO index and leveraging ML models to provide a more robust and objective valuation framework. Furthermore, the implementation of a machine learning model for forecasting post-ICO returns achieves an accuracy of 71%, with the inclusion of social sentiment data and the novel ICO index significantly enhancing its predictive power.

In summary, Chapter 3 contributes to the academic discourse by demonstrating that data-driven token assessments can enhance return prediction in high-risk, low-regulation environments such as ICOs. Moreover, investors also gain new tools to assess risk more accurately, while regulators are reminded of the critical need for transparency and standardised disclosures in whitepapers, as emphasised by Ante et al. (2018). The findings further reveal the impracticality of relying on detailed whitepaper content for predicting token returns, suggesting that investors may primarily treat its presence as a positive signal, regardless of its detailed information (Florysiak and Schandlbauer, 2019; Ante et al., 2018). Regulators can implement investor protections and information disclosure requirements to prevent fraudulent activities, which are highly active in the ICO environment (Hornuf et al., 2022). It is important to acknowledge the limitations of using a lexicon-based sentiment analysis tool like VADER and the inherent biases of relying solely on text published by the ventures themselves, as this content may be crafted to encourage token purchases and influence returns. Future studies can significantly benefit from incorporating data from various perspectives and additional social media platforms, such as Facebook and Reddit. Accessing this broader social media landscape would provide a more comprehensive understanding of public perception towards token issuers, thereby improving the token assessment.

Chapter 4 shifts its focus to the growing, yet underexplored, trend of impact investing in entrepreneurial financing. While research in sustainable finance in public equities clearly demonstrates the power of alternative data and machine learning for informed decision-making (Krappel et al., 2021; Guo et al., 2020; Ruberg et al., 2021; Gutierrez-Bustamante and Espinosa-Leal, 2022), a significant gap exists in the early-stage investment landscape. The absence of established ESG ratings and readily available sustainability reports, primarily due to ventures' limited data disclosure and focus on survivability over transparency, poses a significant challenge for investors. Addressing this gap, the chapter proposes a novel startup valuation framework that leverages machine learning and unstructured text data, including startup news articles and sustainability reports, as a complement to traditional financial statement-based approaches such as discounted cash flow models (Williams, 1938). While acknowledging the importance of traditional factors like founder characteristics, the research reveals a new trend: sustainability considerations are profoundly influencing investor decisions and company valuations, particularly for early-stage startups. By evaluating the semantic similarity between startup news and established sustainability frameworks, the study introduces a predictive model that improve valuation accuracy by 16.45% over a baseline regression model using only startup and funding round characteristics.

These findings contribute to the broader discussion on integrating environmental, social, and governance (ESG) considerations into entrepreneurial financing, in addition to token financing examined by Mansouri and Momtaz (2022). The study supports calls for a more systematic inclusion of sustainability in investment processes, as emphasised by Lin (2022), and highlights the potential for developing dedicated sustainability-focused investment vehicles. From a practical perspective, incorporating sustainability indicators from alternative text sources can help investors make more informed and responsible decisions while fostering trust with limited partners through a demonstrated commitment to long-term value and social goals. The chapter also highlights important policy implications. As sustainability becomes a central factor in investment decision-making, there is a growing need for regulatory frameworks that ensure transparency, prevent greenwashing, and support consistent sustainability reporting standards. These findings align with initiatives such as the EU Sustainable Finance Disclosure Regulation (SFDR) (European Comission, 2022; Roure, 2024) and suggest that similar disclosure mandates may be broadened to venture capital and private equity to promote accountability and sustainable investment practices. In addition, the study can be extended to other sources of sustainability and generate ESG scores for early-stage companies, which are not currently available in the financial markets. Additionally, further studies may explore sentiment dynamics or cross-sector comparisons to better understand how sustainability is perceived and priced in different entrepreneurial ecosystems.

Lastly, Chapter 5 systematically explores a novel application of reinforcement learning (RL) within the context of private capital investment, an area that remains largely underexplored in academic research (Afsar et al., 2022). While RL showcases increasing success in public equity markets (Deng et al., 2017; Azhikodan et al., 2019; Charpentier et al., 2023), its integration into private markets, particularly venture capital, is still in its early stages. This can be explained by the nature of venture capital investment, specifically its illiquidity and highly skewed return profile, which collectively pose a challenge for designing an agent to interact effectively within this environment. This chapter addresses this gap by proposing a reinforcement learning-based recommendation system tailored to the unique characteristics of venture capital investing, including long investment horizons, high uncertainty, and sparse, delayed reward signals. The study designs and evaluates a reinforcement learning model capable of recommending startups with strong potential for successful exits, such as mergers and acquisitions (M&A) or initial public offerings (IPOs). Key components of the RL architecture, including state representations, reward functions, exploration strategies, and hyperparameter configurations, lay the groundwork for future RL applications in illiquid markets. The model demonstrates strong performance by consistently identifying startups associated with higher average funding amounts, which is a critical indicator of future investment potential. Furthermore, the system shows flexibility in supporting both generalised and specialised investment strategies, offering practical value to venture capitalists seeking to improve decision-making during the screening and due diligence processes.

These findings contribute to the research at the intersection of artificial intelligence and financial decision-making. The study highlights the importance of adapting reinforcement learning models to suit the unique dynamics of venture capital, where traditional RL techniques must be refined to handle limited feedback, uncertain outcomes, and evolving investment criteria. The results indicate that RL systems have the potential to enhance decision-making in private capital markets, leading to stronger portfolios and more efficient capital allocation. Despite these encouraging outcomes, the research acknowledges several limitations. The dataset size of 1,000 UK startups not only limits the model's generalisability but also significantly restricts the effective application of more complex deep learning architectures, which typically require large datasets to capture intricate relationships. Future research should include a broader range of startups across different regions and industries. Additionally, exploring alternative reinforcement learning architectures beyond Q-learning, such as policy gradient or actor-critic methods, may offer better performance in larger and more complex venture environments. Finally, the study does not fully account for certain factors, such as systematic risks, that may influence the performance of the recommendations.

This thesis demonstrates the transformative potential of alternative datasets in addressing key challenges in entrepreneurial finance, particularly information asymmetry, by complementing structured data from commercial sources with advanced machine learning methodologies. The findings open several promising avenues for future research. One direction involves extending these approaches to private equity investments in mature companies that do not exit through initial public offerings. Another unexplored area is the application of machine learning models to the post-investment stage, an area beyond this thesis's current scope. As the technological landscape evolves, this research can also pivot toward innovative methodologies such as data augmentation to artificially increase the size and diversity of a training dataset for textual financial data, a domain still nascent in financial markets (Bayer et al., 2022). Additionally, the rapid advancement of generative AI presents exciting opportunities for both academic and practical exploration. For instance, generative models are developed to support startup operations (Tran and Murphy, 2023), enhance document analysis, and improve reasoning processes in financial decision-making, thereby empowering analysts and venture capitalists (Desai et al., 2024).

Strategic and well-informed investment decisions are essential to driving innovation and sustainable growth within the venture capital ecosystem. The influence of venture capital extends beyond financial returns, fostering economic development and generating societal value. By integrating alternative data and advanced machine learning techniques, this thesis contributes to and encourages building more intelligent, data-driven, and responsible investment practices for the future.

Appendix A

Variables description - Chapter 3

Table A.1: A list of ICO variables and descriptions used in this study.

Name	**Description**
Soft Cap	A minimum capital requirement to deliver the product and service in USD.
Whitelist/KYC	An indication of whether the Know-Your-Customer (KYC) and whitelist are performed prior to purchasing the token.
Pre-ICO	An indication of whether the pre-ICO or pre-selling is available.
IEO	An indication of whether the token sale is handled and vetted by an exchange, i.e., Initial exchange offering (IEO).
Bonus	An indication of whether the bonus is available for investors who purchase the token.
ICO Duration	The number of days taken during the ICO process.
Token Listing Duration	The number of days for a token to be listed on the secondary market or crypto exchange.
Token for Sale	The number of tokens available for sale.
% Token Sale	The percentage distribution of tokens available for sale.
Token Type	The type of the token issued on the blockchain such as utility token, ERC-based token.
Platform	The blockchain platform name that the venture issues the token on, such as Ethereum or their own blockchain network.
ETH-based	An indication of whether the application is operated on Ethereum blockchain.
Team Size	The number of team members involved in the blockchain application development, including the management team, developer team, and advisors.
CEO	An indication of whether the venture has a Chief Executive Officer (CEO).

Table A.1 continued from previous page

Name	Description
CTO	An indication of whether the venture has a Chief Technology Officer (CTO).
CEO Prev Experience	An indication of whether a CEO has prior experience of running blockchain projects.
Whitepaper Disclosure	An indication of whether the venture has a whitepaper.
Problem Description Disclosure	A sentiment of problem description aspect on the whitepaper.
Technical Disclosure	A sentiment of technical aspect on the whitepaper.
Roadmap Disclosure	A sentiment of product roadmap aspect on the whitepaper.
Team Disclosure	A sentiment of team aspect on the whitepaper.
Financial Disclosure	A sentiment of token and financial aspect on the whitepaper.
Business Landscape Disclosure	A sentiment of business landscape aspect on the whitepaper.
Risk Disclosure	A sentiment of risk aspect on the whitepaper.
Twitter Activity	The number of tweets posted by the venture published during the ICO.
% Positive Tweets	A ratio of positive sentiments of tweets posted by the venture.
Market Size	The number of markets or exchanges that sell the token.
Country	The country where the token has been issued or launched.
Restricted Area	The number of countries that are restricted for the project to operate.
Social Media	The number of social media platforms used by the venture as communication channels
ICO Rating	The rating of ICO is given by the assessment algorithm and experts on the ICObench platform.
First-day Token Return	A log return of investment on the first day that the token is listed on the exchange.
180 days Token Return	A log return of investment 180 day after the token is listed on the exchange.

Appendix B

Correlation table - Chapter 3

Table B.1: The correlation table of the numerical variable used in the study.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Team Size		0.148	0.093	0.202	-0.032	-0.016	-0.074	-0.079	0.061	0.028	0.050	0.264	-0.078
(2) Social Media	0.148		-0.112	-0.038	0.120	-0.077	0.144	-0.028	0.082	0.086	0.124	0.553	-0.020
(3) Soft Cap (log)	0.093	-0.112		0.223	0.101	-0.118	0.082	0.096	-0.209	-0.048	0.112	-0.006	-0.043
(4) Token for Sale (log)	0.202	-0.038	0.223		0.008	0.053	0.111	-0.004	0.018	-0.075	0.030	0.131	-0.019
(5) % Token Sale	-0.032	0.120	0.101	0.008		-0.065	0.133	0.014	-0.103	0.027	0.120	0.098	-0.041
(6) Market Size	-0.016	-0.077	-0.118	0.053	-0.065		-0.096	-0.040	0.068	-0.006	-0.049	0.077	0.214
(7) ICO Duration	-0.074	0.144	0.082	0.111	0.133	-0.096		0.086	0.148	0.012	0.525	0.122	0.042
(8) Token Listing Duration	-0.079	-0.028	0.096	-0.004	0.014	-0.040	0.086		0.034	-0.039	0.180	0.006	0.099
(9) Restricted Area	0.061	0.082	-0.209	0.018	-0.103	0.068	0.148	0.034		0.011	0.043	0.171	-0.109
(10) % Positive Tweets	0.028	0.086	-0.048	-0.075	0.027	-0.006	0.012	-0.039	0.011		0.015	0.063	0.029
(11) Twitter Activity	0.050	0.124	0.112	0.030	0.120	-0.049	0.525	0.180	0.043	0.015		0.153	0.024
(12) ICO Rating	0.264	0.553	-0.006	0.131	0.098	0.077	0.122	0.006	0.171	0.063	0.153		-0.094
(13) 180 days Token Return (log)	-0.078	-0.020	-0.043	-0.019	-0.041	0.214	0.042	0.099	-0.109	0.029	0.024	-0.094	

Table B.2: The correlation table of the ordinal variable used in the study.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Whitelist/KYC		0.340	0.120	0.155	-0.042	-0.020	-0.076	0.042	0.240	-0.062	0.065	0.168		0.132
(2) Pre-ICO	0.340		0.065	0.075	-0.050	-0.008	0.001	-0.062	0.315	-0.014	0.121	0.037		0.114
(3) CEO	0.120	0.065		0.447	0.091	0.007	0.121	0.022	0.166	-0.060	0.012	0.091		-0.057
(4) CTO	0.155	0.075	0.447		-0.060	0.042	0.076	-0.006	0.116	-0.082	0.207	0.119		0.117
(5) CEO Prev Experience	-0.042	-0.050	0.091	-0.060		0.083	-0.042	-0.047	0.048	0.048	0.044	0.037		-0.572
(6) Whitepaper Disclosure	-0.020	-0.008	0.007	0.042	0.083		0.044	-0.005	-0.041					
(7) ETH-based	-0.076	0.001	0.121	0.076	-0.042	0.044		-0.064	0.030	-0.032	0.200	0.071		0.273
(8) IEO	0.042	-0.062	0.022	-0.006	-0.047	-0.005	-0.064		-0.079	0.009	-0.003	0.066		0.027
(9) Bonus	0.240	0.315	0.166	0.116	0.048	-0.041	0.030	-0.079		-0.121	0.077	0.089		-0.103
(10) Sentiment Problem Description Disclosure	-0.062	-0.014	-0.060	-0.082	0.048		-0.032	0.009	-0.121		0.294	-0.031		-0.020
(11) Sentiment Technical Disclosure	0.065	0.121	0.012	0.207	0.044		0.200	-0.003	0.077	0.294		-0.031		-0.017
(12) Sentiment Roadmap Disclosure	0.168	0.037	0.091	0.119	0.037		0.071	0.066	0.089	-0.031	-0.031			
(13) Sentiment Team Disclosure														
(14) Sentiment Finance Disclosure	0.132	0.114	-0.057	0.117	-0.572		0.273	0.027	-0.103	-0.020	-0.017			
(15) Sentiment Business Landscape Disclosure	-0.138	-0.151	-0.112	-0.036	0.034		0.190	0.049	0.057	0.216	0.328	-0.034		
(16) Sentiment Risk Disclosure	0.094	0.385	0.076	-0.164			-0.081	-0.219	0.219		0.143	0.171		
(17) 180 days Token Return (log)	-0.138	-0.190	-0.105	-0.058	0.019	0.090	-0.043	-0.094	-0.127	-0.034	0.055	0.130		-0.120

	(15)	(16)	(17)
(1) Whitelist/KYC	-0.138	0.094	-0.138
(2) Pre-ICO	-0.151	0.385	-0.190
(3) CEO	-0.112	0.076	-0.105
(4) CTO	-0.036	-0.164	-0.058
(5) CEO Prev Experience	0.034		0.019
(6) Whitepaper Disclosure			0.090
(7) ETH-based	0.190	-0.081	-0.043
(8) IEO	0.049	-0.219	-0.094
(9) Bonus	0.057	0.219	-0.127
(10) Sentiment Problem Description Disclosure	0.216		-0.034
(11) Sentiment Technical Disclosure	0.328	0.143	0.055
(12) Sentiment Roadmap Disclosure	-0.034	0.171	0.130
(13) Sentiment Team Disclosure			
(14) Sentiment Finance Disclosure			-0.120
(15) Sentiment Business Landscape Disclosure			0.047
(16) Sentiment Risk Disclosure			-0.221
(17) 180 days Token Return (log)	0.047	-0.221	

Table B.3: The correlation table of the nominal variable used in the study.

	(1)	(2)	(3)	(4)
(1) Country		0.018	0.047	-0.084
(2) Platform	0.018		0.565	0.008
(3) Token Type	0.047	0.565		0.105
(4) 180 days Token Return (log)	-0.084	0.008	0.105	

Appendix C

Adam algorithm - Chapter 4

Algorithm 1 Adam Algorithm for Stochastic Optimisation used in PyTouch library (Kingma and Ba, 2017; Paszke et al., 2019)

```
    Input: $\gamma, \beta_1, \beta_2, \theta_0, f(\theta), \lambda, amsgrad, maximise$
    Initialise: $m_0 \leftarrow 0, v_0 \leftarrow 0, \hat{v}_t^{max} \leftarrow 0$

  for $t = 1$ do
      if $maximise$ then
          $g_t \leftarrow -\nabla_\theta f_t(\theta_{t-1})$
      else
          $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
      end if
      if $\lambda \neq 0$ then
          $g_t \leftarrow g_t + \lambda\theta_{t-1}$
      end if
      $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)g_t$
      $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)g_t^2$
      $\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$
      $\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$
      if $amsgrad$ then
          $\hat{v}_t \leftarrow max(\hat{v}_t^{max}, \hat{v}_t)$
          $\theta_t \leftarrow \theta_{t-1} - \gamma\hat{m}_t/(\sqrt{\hat{v}_t^{max}+\varepsilon})$
      else
          $\theta_t \leftarrow \theta_{t-1} - \gamma\hat{m}_t/(\sqrt{\hat{v}_t+\varepsilon})$
      end if
  end for
```

Appendix D

Variables description - Chapter 4

Table D.1: Description of Crunchbase variables for startup valuation model.

Variable Name	Type	Description
Organisation		
Country	Nominal	The numerical label of country that the organisation currently operating in or its headquarter^1.
Company Status	Nominal	The numerical label of current status of the organisation^{-1}.
Has Facebook	Binary	Indicator whether the organisation has Facebook social media.
Has LinkedIn	Binary	Indicator whether the organisation has LinkedIn social media.
Has Twitter	Binary	Indicator whether the organisation has Twitter social media.
Employee Count	Ordinal	A numerical label of organisation size grouped into a category.
Company Founded Year	Numerical	A year that organisation has established.
Top Ten City	Binary	Indicator whether the organisation has established in the top ten cities that has the highest number of startups globally.

Table D.1 continued from previous page

Variable Name	Type	Description
Industry Category	Binary	Indicator whether the organisation has been operated in the following industry and model: Administrative Services, Advertising, Agriculture and Farming, Apps, Artificial Intelligence, Biotechnology, Clothing and Apparel, Commerce and Shopping, Community and Lifestyle, Consumer Electronics, Consumer Goods, Content and Publishing, Data and Analytics, Design, Education, Energy, Events, Financial Services, Food and Beverage, Gaming,Government and Military, Hardware, Health Care, Information Technology, Internet Services, Lending and Investments, Manufacturing, Media and Entertainment, Messaging and Telecommunications, Mobile, Music and Audio, Natural Resources, Navigation and Mapping, Other, Payments, Platforms, Privacy and Security, Professional Services, Real Estate, Sales and Marketing, Science and Engineering, Software, Sports, Sustainability, Transportation, Travel and Tourism, Video
Founder and Co-Founder		
Founder Count	Numerical	A number of founders and co-founders of the organisation.
Has Bachelor	Numerical	A number of founders and co-founders hold bachelor’s degrees.
Has Master	Numerical	A number of founders and co-founders hold master’s degrees.

Table D.1 continued from previous page

Variable Name	Type	Description
Has MBA	Numerical	A number of founders and co-founders completed the MBA course.
Has PhD	Numerical	A number of founders and co-founders hold PhD degrees.
Top 100 Education	Numerical	A number of founder and co-founder attended the top 100 universities given the QS global ranking[2].
Top 50 Education	Numerical	A number of founders and co-founders attended the top 50 universities given the QS global ranking[0].
Top 10 Education	Numerical	A number of founders and co-founders attended the top 10 universities given the QS global ranking[0].
STEM Education	Numerical	A number of founders and co-founders hold STEM degrees.
Funding Round		
Investment Type	Nominal	The numerical label of funding series[-1].
Log Amount Raised	Numerical	A log value of amount raised in funding round in USD.
Log Pre-money Valuation	Numerical	A log value of pre-money valuation of the organisation in each funding round in USD.
Deal Announced Year	Numerical	A year that organisation receive funding rounds.
Deal Age	Numerical	A difference between announced year of funding rounds and funded year of organisation.
Number Funding Rounds	Numerical	The number of funding rounds that the organisation has been participated.
Log Cumulative Amount Raised	Numerical	A log value of cumulative amount raised up to the current funding round in USD.
Investor		
Investor Count	Numerical	A number of investors participated in funding round.

Table D.1 continued from previous page

Variable Name	Type	Description
Top Institutional Investor	Binary	Indicator whether top ten institutional investors (e.g., venture capital fund, private equity fund) participated in this funding round.
Top Individual Investor	Binary	Indicator whether the top ten individual investors participated in this funding round.
Investment Age	Numerical	A difference between announced year of funding rounds and funded year of organisation.
Accelerator Investor	Binary	Indicator whether the top investors participated in funding round are incubator or accelerator.
Angle Investor	Binary	Indicator whether the top investors participated in funding round are angle investor.
Different Geolocation	Binary	Indicator whether the country of operation of startups and investors is different.
Log Cumulative Investment Amount	Numerical	A log value of the cumulative amount raised that investors have participated in up to the current funding round in USD.
Investor Experience	Numerical	A number of funding rounds that investors have participated in up to the announced year of the funding round.

Appendix E

Correlation - Chapter 4

Table E.1: Correlation matrix of Crunchbase variables and startup valuation.

	Log Pre-Money Valuation
Country	-0.152
Company Status	0.071
Has Facebook	0.025
Has LinkedIn	0.066
Has Twitter	-0.046
Employee Count	0.720
Company Founded Year	-0.223
Top Ten City	0.002
Founder Count	0.438
Has Bachelor	0.393
Has Master	0.357
Has MBA	0.344
Has PhD	0.230
Top 100 Education	0.395
Top 50 Education	0.388
Top 10 Education	0.376
STEM Education	0.373
Investment Type	0.488
Log Amount Raised	0.860
Deal Announced Year	0.269
Deal Age	0.453
Number Funding Rounds	0.508
Log Cumulative Amount Raised	0.903
Investor Count	0.163
Top Institutional Investor	0.166
Top Individual Investor	0.029
Investment Age	0.199

Accelerator Investor	-0.007
Angle Investor	-0.154
Different Geolocation	0.015
Log Cumulative Investment Amount	0.100
Investor Experience	0.129

Appendix F

News and sustainability text similarity - Chapter 4

Table F.1: Startup News with Similarity to Sustainability Frameworks.

Topics	Average Similarity	Example of Top Three News Title
UN SDGs		
No Poverty	0.409	- Providing healthcare to lower-income communities values Cityblock Health at $1 billion, - Apeel gets more cash to fight poverty and food insecurity in emerging markets with its food-preserving tech, - Investment tech won't solve systemic wealth gaps, but it's a good start
Zero Hunger	0.492	- Apeel gets more cash to fight poverty and food insecurity in emerging markets with its food-preserving tech, - Preventing food waste nets Apeel $250 million from Singapore's government, Oprah and Katy Perry, - Farmers Business Network raises $20 million to help farmers avoid spending on what they don't need
Good Health and Wellbeing	0.460	- Providing healthcare to lower-income communities values Cityblock Health at $1 billion, - Quizlet valued at $1 billion as it raises millions during a global pandemic, - Oscar Health's CEO believes the U.S. has a moral obligation to provide healthcare to its citizens
Quality Education	0.392	- India's Vedantu scores $24M more for its online tutoring service, - TikTok makes education push in India, - Education technology meets its limits

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Gender Equality	0.403	- Are option grants promoting gender and racial inequity?, - Tech’s new diversity leaders explain how they plan to fix sexism and racism in the industry, - Female Founders: The State Of The Union
Clean Water and Sanitation	0.390	- Apeel gets more cash to fight poverty and food insecurity in emerging markets with its food-preserving tech, - Focusing on human and climate health, S2G Ventures launches ocean fund with $100 million in commitments, - Preventing food waste nets Apeel $250 million from Singapore’s government, Oprah and Katy Perry
Affordable and Clean Energy	0.436	- 4 sustainable industries where founders and VCs can see green by going green, - Southeast Asia’s Grab plans electric vehicle push, - 5G promises to transform the world again
Decent Work and Economic Growth	0.434	- Human Capital: Moving away from master/slave; terminology, - Investment tech won’t solve systemic wealth gaps, but it’s a good start, - Politicized “Gig Economy”; May Make Changing Status Quo More Difficult
Industry Innovation and Infrastructure	0.518	- Latin America’s digital transformation is making up for lost time, - Investment tech won’t solve systemic wealth gaps, but it’s a good start, - Google report: Southeast Asia’s digital economy to triple to $240 billion by 2025

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Reduced Inequalities	0.435	- Investment tech won't solve systemic wealth gaps, but it's a good start, - Getaround, Facebook, AI chips, Nvidia, Africa, and immigration, - Developing a global financial architecture
Sustainable Cities and Communities	0.448	- Startups Weekly: How will we build the city of the future?, - With Detroit Taking A Lyft In A Driverless Car, What's Next For Cities?, - Ride Sharing Will Give Us Back Our Cities
Responsible Consumption and Production	0.504	- Apeel gets more cash to fight poverty and food insecurity in emerging markets with its food-preserving tech, - 4 sustainable industries where founders and VCs can see green by going green, - Preventing food waste nets Apeel $250 million from Singapore's government, Oprah and Katy Perry
Climate Action	0.398	- Stripe Climate is a new tool to let Stripe customers make carbon removal purchases, - Focusing on human and climate health, S2G Ventures launches ocean fund with $100 million in commitments, - Could lessons from the challenger bank revolution kick-start innovation on the climate crisis?

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Life Below Water	0.406	- Stripe Climate is a new tool to let Stripe customers make carbon removal purchases, - Focusing on human and climate health, S2G Ventures launches ocean fund with $100 million in commitments, - Preventing food waste nets Apeel $250 million from Singapore's government, Oprah and Katy Perry
Life On Land	0.349	- Stripe Climate is a new tool to let Stripe customers make carbon removal purchases, - 4 sustainable industries where founders and VCs can see green by going green, - Preventing food waste nets Apeel $250 million from Singapore's government, Oprah and Katy Perry
Peace Justice and Strong Institutions	0.324	- A look ahead at blockchain's next decade, - The Khashoggi murder isn't stopping SoftBank's Vision Fund, - Announcing the TechCrunch Session on Blockchain Agenda
Partnerships For The Goals	0.435	- Apeel gets more cash to fight poverty and food insecurity in emerging markets with its food-preserving tech, - Developing a global financial architecture, - Latin America's chronic inefficiency could drive more O2O commerce growth
SASB 26 General Issues		

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Product Quality and Safety	0.417	- The hidden cost of food delivery, - 14 wildly hot takes we need on this Whole Foods + Amazon thing, - Blue Apron Delivers All The Ingredients You Need To Cook Fresh Meals Every Week
Data Security	0.554	- DataGuard, which provides GDPR and privacy compliance-as-a-service, raises $20M, - Why commerce companies are the advertising players to watch in a privacy-centric world, - The Internet Giveth, And Taketh Away: Sometimes, Business Decisions Are Bad For Users
Customer Welfare	0.470	- Assessing the potential for a gig economy in education, - Lucid Lane has developed a service to get patients off of pain meds and avoid dependence, - Crimson Education, a platform to help students get into top universities, nabs $5M at a $245M valuation
Energy Management	0.464	- 4 sustainable industries where founders and VCs can see green by going green, - Hewlett Packard Enterprise Places A Big Bet On Containers, - A Look At Startup Opportunities In The Container Era
Employee Engagement, Diversity and Inclusion	0.504	- Human Capital: Google's labor stumbles, - Placement is the much-needed talent agent for jobseekers, - Tech's new diversity leaders explain how they plan to fix sexism and racism in the industry

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Product Design and Lifecycle Management	0.477	- Could lessons from the challenger bank revolution kick-start innovation on the climate crisis?, - How digital has redefined go-to-market strategy, - The Power Of Online-To-Offline Is Moving Beyond Local Commerce
Customer Privacy	0.620	- BigID bags another $50M round as data privacy laws proliferate, - Why commerce companies are the advertising players to watch in a privacy-centric world, - The Internet Giveth, And Taketh Away: Sometimes, Business Decisions Are Bad For Users
Employee Health and Safety	0.419	- The two forces reshaping the landscape of shipping and logistics, - Lyft is getting more serious about autonomous vehicle safety with new hire, - Transport’s coming upheaval
Materials Sourcing and Efficiency	0.472	- Amazon’s next conquest will be apparel, - How legacy brands and retailers can keep up with our tech-driven world, - Farmers Business Network raises $20 million to help farmers avoid spending on what they don’t need
Supply Chain Management	0.435	- For alternative meat manufacturer Beyond Meat, fast food chains giveth and take away, - The Amazonization of Whole Foods, one year in, - 14 wildly hot takes we need on this Whole Foods + Amazon thing

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Access and Affordability	0.496	- The tale of 2 challenger bank models, - Fundera, Funding Circle And Others Introduce The Small Business Borrowers' Bill Of Rights, - Startup Financial Services Companies Come Of Age
Selling Practices and Product Labeling	0.501	- Media roundup: Google to cut big checks for news publishers, Substack continues to draw top creators, more, - Now more than ever we need fintechs to lead on consumer transparency, - Crimson Education, a platform to help students get into top universities, nabs $5M at a $245M valuation
Human Rights and Community Relations	0.477	- Synthetic biology startups are giving investors an appetite, - Benchling's software for managing biotech research nabs $34.5 million, - Science37 aims to democratize clinical research with a fresh $29 million in growth funding
Business Ethics	0.493	- Now more than ever we need fintechs to lead on consumer transparency, - The two forces reshaping the landscape of shipping and logistics, - Investors are pouring money into Latin America's logistics and shipping businesses

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Competitive Behavior	0.507	- Indian startups explore alliance and alternative app store to fight Google's 'monopoly', - Trolling the patent trolls, - The Internet Giveth, And Taketh Away: Sometimes, Business Decisions Are Bad For Users
Systemic Risk Management	0.595	- Enterprise companies find MLOps critical for reliability and performance, - Startups are helping cloud infrastructure customers avoid vendor lock-in, - Google's Cloud outage is resolved, but it reveals the holes in cloud computing's atmosphere
Waste and Hazardous Materials Management	0.539	- A glint of hope for India's food delivery market as Zomato projects monthly cash burn of less than $1M, - Preventing food waste nets Apeel $250 million from Singapore's government, Oprah and Katy Perry, - The hidden cost of food delivery
Critical Incident Risk Management	0.467	- Uber commits $50 million to safety supplies for drivers, - Investigation finds e-scooters a cause of 1,500+ accidents, - Lyft is getting more serious about autonomous vehicle safety with new hire
GHG Emissions	0.500	-Lyft invests millions of dollars to offset its effect on climate change, - Regulators order environmental impact study of Lyft Line and UberPOOL, - Europe's DocPlanner Bags $10M To Grow Its Healthcare Booking Platform

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Ecological Impacts	0.374	- The two forces reshaping the landscape of shipping and logistics, - Regulators order environmental impact study of Lyft Line and UberPOOL, - HotelTonight ExpandsTo The UK, Former Jetsetter Leads The Charge
Air Quality	0.443	- Air quality monitoring service Airly raises $2 million as fires, pollution force consumers to take note, - Regulators order environmental impact study of Lyft Line and UberPOOL, - Regulators Should Favor Lyft And Uber, Not Taxis For Safety Reasons
Water and Wastewater Management	0.422	- 4 sustainable industries where founders and VCs can see green by going green, - Farmers Business Network raises $20 million to help farmers avoid spending on what they don’t need, - Barry Sternlicht, Former CEO Of Hotel Giant Starwood, Invests In HotelTonight
Labor Practices	0.567	- Equity Monday: Food delivery economics, and global layoffs, - Instacart shoppers plan a series of actions in protest of company’s wage practices, - HotelTonight Cuts 20 Percent Of Its Workforce

Table F.1 continued from previous page

Topics	**Average Similarity**	**Example of Top Three News Title**
Physical Impacts of Climate Change	0.444	- Cowboy VC's Aileen Lee: Your coronavirus scenario planning should be more conservative, - Could lessons from the challenger bank revolution kick-start innovation on the climate crisis?, - VC doors are wide open for real estate startups
Management of the Legal and Regulatory Environment	0.186	- India's Ather Energy raises $51 million to grow its electric scooters business, - Electric scooter maker Gogoro raises $300 million for growth, - VCs are betting on the great Chinese fitness boom
Business Model Resilience	0.274	- India's Ather Energy raises $51 million to grow its electric scooters business, - Quick-charging battery startup StoreDot gets $60M on $500M valuation led by Daimler, - Gogoro's Compact New Electric Scooter Charging Stations Can Be Installed Inside Homes And Stores
SASB 5 Disclosure Dimensions		
Environment	0.438	- 4 sustainable industries where founders and VCs can see green by going green, - Hewlett Packard Enterprise Places A Big Bet On Containers, - A Look At Startup Opportunities In The Container Era

Table F.1 continued from previous page

Topics	Average Similarity	Example of Top Three News Title
Social Capital	0.521	- DataGuard, which provides GDPR and privacy compliance-as-a-service, raises $20M, - BigID bags another $50M round as data privacy laws proliferate, - Segment's new privacy portal helps companies comply with expanding regulations
Human Capital	0.491	- Placement is the much-needed talent agent for jobseekers, - Tech's new diversity leaders explain how they plan to fix sexism and racism in the industry, - An Open Letter To Those Not Employed At Instagram
Business Model	0.462	- How legacy brands and retailers can keep up with our tech-driven world, - Narvar raises $22 million to help internet retailers deliver physical goods without frustrating customers, - The Power Of Online-To-Offline Is Moving Beyond Local Commerce
Leadership and Governance	0.475	- Startups are helping cloud infrastructure customers avoid vendor lock-in, - Trolling the patent trolls, - The Internet Giveth, And Taketh Away: Sometimes, Business Decisions Are Bad For Users

Appendix G

SASB general issue disclosure - Chapter 4

Table G.1: A mapping of SASB disclosure topics to the high-level dimension (IFRS Foundation, 2022a). An issue category with asterisk (*) means that ESG-BERT has additional category called Director Removal.

Dimension	General Issue Category	Availability in ESG-BERT
Environment	GHG Emissions	Y
	Air Quality	Y
	Energy Management	Y
	Water & Wastewater Management	Y
	Waste & Hazardous Materials Management	Y
	Ecological Impacts	Y
Social Capital	Human Rights & Community Relations	Y
	Customer Privacy	Y
	Data Security	Y
	Access & Affordability	Y
	Product Quality & Safety	Y
	Customer Welfare	Y
	Selling Practices & Product Labeling	Y
Human Capital	Labor Practices	Y
	Employee Health & Safety	Y
	Employee Engagement, Diversity & Inclusion	Y
Business Model and Innovation	Product Design & Lifecycle Management	Y
	Business Model Resilience	Y
	Supply Chain Management	Y

Table G.1 continued from previous page

Dimension	General Issue Category	Availability in ESG-BERT
	Materials Sourcing & Efficiency	N*
	Physical Impacts of Climate Change	Y
Leadership and Governance	Business Ethics	Y
	Competitive Behavior	Y
	Management of the Legal & Regulatory Environment	Y
	Critical Incident Risk Management	Y
	Systemic Risk Management	Y

Appendix H

Example of Refinitiv’s ESG Disclosure - Chapter 4

Table H.1: Example of ESG disclosure structure from the Refinitiv Workspace (Refinitiv, 2022).

Dimension	Example Disclosure	Measurement
Environment	Does the company make use of Renewable energy?	True/False
	Total CO2 and CO2 equivalents emission	in Tons
Social	Does the company have a policy to improve employee health and safety within the company and its supply chain?	True/False
	Percentage of women managers among total managers in the company	in percentage
Governance	Does the company have an audit board committee?	True/False
	Ratio of CEO's total compensation over median employee compensation as reported by the company	in ratio

[-1] Transformed by LabelEncoder available in scikit-learn library.

[0] The list of universities obtained from https://www.topuniversities.com/university-rankings/world-university-rankings/2021 accessed on 12 July 2023.

Appendix I

Model parameters - Chapter 4

Table I.1: Hyperparameters of regression models.

Model	R-Squared	MSE	Parameters
Crunchbase			
LR	0.747	0.617	-
RF	0.691	0.752	{’n_estimators’: 200, ’min_samples_split’: 10, ’min_samples_leaf’: 2, ’max_features’: ’sqrt’, ’max_depth’: 10, ’bootstrap’: False}
GB	0.743	0.626	{’subsample’: 1.0, ’n_estimators’: 100, ’min_samples_split’: 4, ’max_depth’: 3, ’learning_rate’: 0.1}
DL	0.541	1.118	{’num_epochs’: 100, ’learning_rate’: 0.001}
TechCrunch			
Doc2Vec + LR	-0.398	3.484	{’doc2vec_vector_size’: 50}
Doc2Vec + RF	-0.157	2.884	{’n_estimators’: 100, ’min_samples_split’: 10, ’min_samples_leaf’: 4, ’max_features’: ’sqrt’, ’max_depth’: 5, ’bootstrap’: True, ’doc2vec_vector_size’: 50}
Doc2Vec + GB	-0.039	2.589	{’subsample’: 1.0, ’n_estimators’: 200, ’min_samples_split’: 6, ’max_depth’: 3, ’learning_rate’: 0.001, ’doc2vec_vector_size’: 100}
Doc2Vec + DL	-0.513	3.771	{’num_epochs’: 50, ’learning_rate’: 0.0001}
BERT + LR	-0.486	3.704	
BERT + RF	0.007	2.475	{’n_estimators’: 300, ’min_samples_split’: 5, ’min_samples_leaf’: 4, ’max_features’: ’sqrt’, ’max_depth’: 5, ’bootstrap’: True}
BERT + GB	-0.021	2.545	{’subsample’: 0.8, ’n_estimators’: 200, ’min_samples_split’: 2, ’max_depth’: 7, ’learning_rate’: 0.001}

Table I.1 continued from previous page

Model	R-Squared	MSE	Parameters
BERT + DL	-0.235	3.078	{'num_epochs': 50, 'learning_rate': 0.0001}
FinBERT + LR	-0.352	3.369	-
FinBERT + RF	-0.017	2.536	{'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 5, 'bootstrap': True}
FinBERT + GB	-0.024	2.553	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 2, 'max_depth': 5, 'learning_rate': 0.001}
FinBERT + DL	-0.178	2.937	{'num_epochs': 50, 'learning_rate': 0.0001}
ESG-BERT + LR	-0.431	3.567	-
ESG-BERT + RF	-0.022	2.548	{'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 5, 'bootstrap': True}
ESG-BERT + GB	-0.025	2.555	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 4, 'max_depth': 5, 'learning_rate': 0.001}
ESG-BERT + DL	-0.423	3.547	{'num_epochs': 10, 'learning_rate': 0.001}
Crunchbase & TechCrunch			
Doc2Vec + LR	0.754	0.613	{'doc2vec_vector_size': 150}
Doc2Vec + RF	0.736	0.658	{'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': None, 'bootstrap': True, 'doc2vec_vector_size': 100}
Doc2Vec + GB	0.761	0.595	{'subsample': 1.0, 'n_estimators': 300, 'min_samples_split': 6, 'max_depth': 5, 'learning_rate': 0.1, 'doc2vec_vector_size': 150}
Doc2Vec + DL	0.725	0.687	{'num_epochs': 50, 'learning_rate': 0.001}

Table I.1 continued from previous page

Model	R-Squared	MSE	Parameters
BERT + LR	0.727	0.681	-
BERT + RF	0.432	1.416	{'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}
BERT + GB	0.219	1.946	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 4, 'max_depth': 7, 'learning_rate': 0.001}
BERT + DL	0.667	0.829	{'num_epochs': 50, 'learning_rate': 0.001}
FinBERT + LR	0.706	0.732	-
FinBERT + RF	0.434	1.410	{'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}
FinBERT + GB	0.219	1.947	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 2, 'max_depth': 7, 'learning_rate': 0.001}
FinBERT + DL	0.721	0.696	{'num_epochs': 100, 'learning_rate': 0.001}
ESG-BERT + LR	0.676	0.807	-
ESG-BERT + RF	0.427	1.428	{'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}
ESG-BERT + GB	0.219	1.946	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 2, 'max_depth': 7, 'learning_rate': 0.001}
ESG-BERT + DL	0.640	0.897	{'num_epochs': 50, 'learning_rate': 0.001}
Crunchbase & UN SDG			
LR	0.690	0.755	-

Table I.1 continued from previous page

Model	R-Squared	MSE	Parameters
RF	0.785	0.523	{'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True}
GB	0.770	0.561	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 4, 'max_depth': 3, 'learning_rate': 0.1}
DL	0.456	1.327	{'num_epochs': 100, 'learning_rate': 0.0001}
Crunchbase & SASB General Issue			
LR	0.690	0.755	-
RF	0.785	0.523	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': None, 'bootstrap': True}
GB	0.770	0.561	{'subsample': 0.8, 'n_estimators': 100, 'min_samples_split': 2, 'max_depth': 5, 'learning_rate': 0.1}
DL	0.456	1.327	{'num_epochs': 50, 'learning_rate': 0.001}
Crunchbase & SASB Dimension			
LR	0.746	0.618	-
RF	0.780	0.536	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True}
GB	0.715	0.695	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 4, 'max_depth': 3, 'learning_rate': 0.1}
DL	0.532	1.140	{'num_epochs': 50, 'learning_rate': 0.001}
Crunchbase & Refinitiv			
LR	0.747	0.617	-

Table I.1 continued from previous page

Model	R-Squared	MSE	Parameters
RF	0.772	0.556	{'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': None, 'bootstrap': True}
GB	0.723	0.676	{'subsample': 0.8, 'n_estimators': 200, 'min_samples_split': 6, 'max_depth': 3, 'learning_rate': 0.01}
DL	0.543	1.115	{'num_epochs': 100, 'learning_rate': 0.001}

Appendix J

Deep Q-learning implementation - Chapter 5

Algorithm 2 Deep Q-learning with experience reply implementation previously applied on seven Atari 2600 games (Mnih et al., 2013).

```
Initialise replay memory $D$ to capacity $N$
Initialise action-value function $Q$ with random weights $\theta$
Initialise target action-value function $\hat{Q}$ with weights $\theta^- = \theta$
**for** episode $1, M$ **do** Initialise sequence $s_1 = \{x_1\}$ and preprocess sequence $\phi_1 = \phi(s_1)$
    **for** $t = 1, T$ **do**
        With probability $\epsilon$ select a random action $a_t$
        otherwise select $a_t = argmax_a Q(\phi(s_t), a; \theta)$
        Execute action $a_t$ in the emulator and observe reward $r_t$ and image $x_{t+1}$
        Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
        Store experience $(\phi_t, a_t, r_t, \phi_{t+1})$ in $D$
        Sample random minibatch of experience $(\phi_j, a_j, r_j, \phi_{j+1})$ from $D$
        **if** episode terminates at step $j + 1$ **then**
            $y_j = r_j$
        **else**
            $y_j = r_j + \gamma max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$
        **end if**
        Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ with respect to the weights $\theta$
        Every $C$ steps reset $\hat{Q} = Q$
    **end for**
**end for**
```

Bibliography

Adhami, Saman, Giancarlo Giudici, Stefano Martinazzi. 2018. Why do businesses go crypto? an empirical analysis of initial coin offerings. *Journal of Economics and Business* **100** 64–75. doi: 10.1016/j.jeconbus.2018.04.001. FinTech – Impact on Consumers, Banking and Regulatory Policy.

Afful-Dadzie, Eric, Zuzana Komínková Oplatková, Stephen Nabareseh. 2015. Selecting start-up businesses in a public venture capital financing using fuzzy promethee. *Procedia Computer Science* **60** 63–72. doi: 10.1016/j.procs.2015.08.105. Knowledge-Based and Intelligent Information Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

Afsar, M. Mehdi, Trafford Crump, Behrouz Far. 2022. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys* **55**(7). doi: 10.1145/3543846.

Ahlers, Gerrit KC, Douglas Cumming, Christina Günther, Denis Schweizer. 2015. Signaling in equity crowdfunding. *Entrepreneurship theory and practice* **39**(4) 955–980. doi: 10.1111/etap.12157.

Akerlof, George A. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* **84**(3) 488–500. doi: 10.2307/1879431.

Amsden, Ryan, Denis Schweizer. 2018. Are blockchain crowdsales the new’gold rush’? success determinants of initial coin offerings. doi: 10.2139/ssrn.3163856.

Andrieu, Guillaume, Aurélie Sannajust. 2023. Icos after the decline: A literature review and recommendations for a sustainable development. *Venture Capital* 1–19doi: 10.1080/13691066.2023.2240024.

Andrés, Pablo de, David Arroyo, Ricardo Correia, Alvaro Rezola. 2022. Challenges of the market for initial coin offerings. *International Review of Financial Analysis* **79** 101966. doi: 10.1016/j.irfa.2021.101966.

Ang, Yu Qian, Andrew Chia, Soroush Saghafian. 2022. *Using Machine Learning to Demystify Startups' Funding, Post-Money Valuation, and Success*. Springer International Publishing, Cham, 271–296. doi: 10.1007/978-3-030-75729-8_10.

Ante, Lennart, Philipp Sandner, Ingo Fiedler. 2018. Blockchain-based icos: Pure hype or the dawn of a new era of startup financing? *Journal of Risk and Financial Management* **11**(4). doi: 10.3390/jrfm11040080.

Antretter, Torben, Ivo Blohm, Dietmar Grichnik, Joakim Wincent. 2019. Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights* **11** e00109. doi: 10.1016/j.jbvi.2018.e00109.

Aouni, Belaïd, Cinzia Colapinto, Davide La Torre. 2013. A cardinality constrained stochastic goal programming model with satisfaction functions for venture capital investment decision making. *Annals of Operations Research* **205**(1) 77–88. doi: 10.1007/s10479-012-1168-4.

Araci, Dogu. 2019. Finbert: Financial sentiment analysis with pre-trained language models. Master's thesis, University of Amsterdam. doi: 10.48550/arXiv.1908.10063.

Arroyo, Javier, Francesco Corea, Guillermo Jimenez-Diaz, Juan A. Recio-Garcia. 2019. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access* **7** 124233–124243. doi: 10.1109/ACCESS.2019.2938659.

Arthur, W Brian. 1989. Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal* **99**(394) 116–131. doi: 10.2307/2234208.

Azhikodan, Akhil Raj, Anvitha G. K. Bhat, Mamatha V. Jadhav. 2019. Stock trading bot using deep reinforcement learning. H. S. Saini, Rishi Sayal, A. Govardhan, Rajkumar Buyya, eds., *Innovations in Computer Science and Engineering*. Springer, 41–49. doi: 10.1007/978-981-10-8201-6_5.

Bayar, Onur, Emre Kesici. 2024. The impact of social media on venture capital financing: evidence from twitter interactions. *Review of Quantitative Finance and Accounting* **62**(1) 195–224. doi: 10.1007/s11156-023-01199-4.

Bayer, Markus, Marc-André Kaufhold, Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys* **55**(7) 1–39. doi: 10.1145/3544558.

Bellavitis, Cristiano, Christian Fisch, Johan Wiklund. 2021. A comprehensive review of the global development of initial coin offerings (icos) and their regulation. *Journal of Business Venturing Insights* **15** e00213. doi: 10.1016/j.jbvi.2020.e00213.

Belleflamme, Paul, Thomas Lambert, Armin Schwienbacher. 2014. Crowdfunding: Tapping the right crowd. *Journal of Business Venturing* **29**(5) 585–609. doi: 10.1016/j.jbusvent.2013.07.003.

Benedetti, Hugo, Leonard Kostovetsky. 2021. Digital tulips? returns to investors in initial coin offerings. *Journal of Corporate Finance* **66** 101786. doi: 10.1016/j.jcorpfin.2020.101786.

Berg, Florian, Julian F Kölbel, Roberto Rigobon. 2022. Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance* **26**(6) 1315–1344. doi: 10.1093/rof/rfac033.

Bhat, Harish S., Daniel Zaelit. 2011. Predicting private company exits using qualitative data. Joshua Zhexue Huang, Longbing Cao, Jaideep Srivastava, eds., *Advances* *in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, Berlin, Heidelberg, 399–410. doi: 10.1007/978-3-642-20841-6_33.

Bianchini, R., A. Croce. 2022. The role of environmental policies in promoting venture capital investments in cleantech companies. *Review of Corporate Finance* **2**(3) 587–616. doi: 10.1561/114.00000024.

Binance Academy. 2020. What is an initial exchange offering (ieo)? `https://academy.binance.com/en/articles/what-is-an-initial-exchange-offering-ieo`.

Black, Ervin. 2003. Usefulness of financial statement components in valuation: An examination of start-up and growth firms. *Venture Capital: An International Journal of Entrepreneurial Finance* **5** 47–69. doi: 10.1080/1369106032000006722.

Block, Joern, Massimo Colombo, Douglas Cumming, Silvio Vismara. 2018. New players in entrepreneurial finance and why they are there. *Small Business Economics* **50**. doi: 10.1007/s11187-016-9826-6.

Block, Joern, Geertjan De Vries, Philipp Sandner. 2012. Venture Capital and the Financial Crisis: An Empirical Study across Industries and Countries. *The Oxford Handbook of Venture Capital*. Oxford University Press. doi: 10.1093/oxfordhb/9780195391596.013.0003.

BNP. 2019. The esg global survey 2019. Tech. rep.

Boffo, R., R. Patalano. 2020. Esg investing: Practices, progress and challenges. `www.oecd.org/finance/ESG-Investing-Practices-Progress-and-Challenges.pdf`.

Bordino, Ilaria, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, Ingmar Weber. 2012. Web search queries can predict stock market volumes. *PLOS ONE* **7**(7) 1–17. doi: 10.1371/journal.pone.0040014.

Bourveau, Thomas, Emmanuel T. De George, Atif Ellahie, Daniele Macciocchi. 2022. The role of disclosure and information intermediaries in an unregulated capital market: Evidence from initial coin offerings. *Journal of Accounting Research* **60**(1) 129–167. doi: 10.1111/1475-679X.12404.

Breiman, Leo. 2001. Random forests. *Machine Learning* **45**(1) 5–32. doi: 10.1023/A:1010933404324.

Burns, Lauren, Andrea Moro. 2018. What makes an ico successful? an investigation of the role of ico characteristics, team quality and market sentiment. *SSRN Electronic Journal* doi: 10.2139/SSRN.3256512.

Bygrave, William D. 1987. Syndicated investments by venture capital firms: A networking perspective. *Journal of Business Venturing* **2**(2) 139–154. doi: https://doi.org/10.1016/0883-9026(87)90004-8.

Bygrave, William D. 1988. The structure of the investment networks of venture capital firms. *Journal of Business Venturing* **3**(2) 137–157. doi: https://doi.org/10.1016/0883-9026(88)90023-7.

Bühler, Hans, Lukas Gonon, Josef Teichmann, Ben Wood. 2018. Deep hedging. doi: 10.48550/ARXIV.1802.03042.

Calic, Goran, Elaine Mosakowski. 2016. Kicking off social entrepreneurship: How a sustainability orientation influences crowdfunding success. *Journal of Management Studies* **53**(5) 738–767. doi: https://doi.org/10.1111/joms.12201.

Calvino, Flavio, Chiara Criscuolo, Carlo Menon. 2016. No country for young firms? (29). doi: 10.1787/5jm22p40c8mw-en.

Campino, José, Ana Brochado, Álvaro Rosa. 2022. Initial coin offerings (icos): Why do they succeed? *Financial Innovation* **8**(1) 17. doi: 10.1186/s40854-021-00317-2.

Caragea, Doina, Mark Chen, Theodor Cojoianu, Mihai Dobri, Kyle Glandt, George Mihaila. 2020. Identifying fintech innovations using bert. *2020 IEEE International Conference on Big Data (Big Data)*. 1117–1126. doi: 10.1109/BigData50022.2020.9378169.

Catalini, Christian, Joshua S Gans. 2018. Initial coin offerings and the value of crypto tokens. Working Paper 24418, National Bureau of Economic Research. doi: 10.3386/w24418.

CB Insights updates. 2020. Our first acquisition: Cb insights acquires venturesource data from dow jones. `https://www.cbinsights.com/research/team-blog/dow-jones-venturesource-valuations/`.

Chan, C.S. Richard, Charuta Pethe, Steven Skiena. 2021. Natural language processing versus rule-based text analysis: Comparing bert score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights* **16** e00276. doi: 10.1016/j.jbvi.2021.e00276.

Charpentier, Arthur, Romuald Élie, Carl Remlinger. 2023. Reinforcement learning in economics and finance. *Computational Economics* **62**(1) 425–462. doi: 10.1007/s10614-021-10119-4.

Chawla, N. V., K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16** 321–357. doi: 10.1613/jair.953.

Chen, Kun. 2019. Information asymmetry in initial coin offerings (icos): Investigating the effects of multiple channel signals. *Electronic Commerce Research and Applications* **36** 100858. doi: 10.1016/j.elerap.2019.100858.

Chen, Minmin, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, Ed Chi. 2019. Top-k off-policy correction for a reinforce recommender system. 456–464. doi: 10.1145/3289600.3290999.

Cochrane, John H. 2005. The risk and return of venture capital. *Journal of Financial Economics* **75**(1) 3–52. doi: 10.1016/j.jfineco.2004.03.006.

Coingecko. 2022. Methodology. `https://www.coingecko.com/en/methodology`. Accessed March 30, 2022.

CoinMarketCap. 2022. Bitcoin price today, BTC to USD live, marketcap and chart | CoinMarketCap. `https://coinmarketcap.com/currencies/bitcoin/`.

Cong, Lin William, Beibei Li, Qingquan Tony Zhang. 2021. Alternative data in fintech and business intelligence. *The Palgrave handbook of FinTech and blockchain* 217–242.

Cong, Lin William, Ye Li, Neng Wang. 2020. Tokenomics: Dynamic Adoption and Valuation. *The Review of Financial Studies* **34**(3) 1105–1155. doi: 10.1093/rfs/hhaa089.

Corchado, J., C. Fyfe, B. Lees. 1998. Unsupervised learning for financial forecasting. *Proceedings of the IEEE/IAFE/INFORMS 1998 Conference on Computational Intelligence for Financial Engineering (CIFEr) (Cat. No.98TH8367)*. 259–263. doi: 10.1109/CIFER.1998.690316.

Corea, Francesco, Giorgio Bertinetti, Enrico Maria Cervellati. 2021a. Hacking the venture industry: An early-stage startups investment framework for data-driven investors. *Machine Learning with Applications* **5** 100062. doi: https://doi.org/10.1016/j.mlwa.2021.100062.

Corea, Francesco, Giorgio Bertinetti, Enrico Maria Cervellati. 2021b. Hacking the venture industry: An early-stage startups investment framework for data-driven investors. *Machine Learning with Applications* **5** 100062. doi: 10.1016/j.mlwa.2021.100062.

Cumming, Douglas, Satish Kumar, Weng Marc Lim, Nitesh Pandey. 2022. Mapping the venture capital and private equity research: a bibliometric review and future research agenda. *Small Business Economics* **61** 1–49. doi: 10.1007/s11187-022-00684-9.

Damodaran, Aswath. 2009. Valuing young, start-up and growth companies: Estimation issues and valuation challenges. *SSRN Electronic Journal* doi: 10.2139/ssrn.1418687.

Davydiuk, Tetiana, Deeksha Gupta, Samuel Rosen. 2023. De-Crypto-ing Signals in Initial Coin Offerings: Evidence of Rational Token Retention. *Management Science* **69**(11) 6584–6624. doi: 10.1287/mnsc.2022.4631.

Dealroom. 2023. Healthtech. `https://dealroom.co/guides/healthtech-guide`.

Dealroom. 2024a. Climate tech. `https://dealroom.co/guides/climate-tech`.

Dealroom. 2024b. The state of global vc. URL `https://dealroom.co/guides/global`.

Dellermann, Dominik, Nikolaus Lipusch, Philipp Ebel, Karl Popp, Jan Marco Leimeister. 2017. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *SSRN Electronic Journal* doi: 10.2139/ssrn.3159123.

Deng, Xin, Yen Teik Lee, Zhengting Zhong. 2018. Decrypting coin winners: Disclosure quality, governance mechanism and team networks. *SSRN Electronic Journal* doi: 10.2139/ssrn.3247741.

Deng, Yue, Feng Bao, Youyong Kong, Zhiquan Ren, Qionghai Dai. 2017. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems* **28**(3) 653–664. doi: 10.1109/TNNLS.2016.2522401.

Denis, David J. 2004. Entrepreneurial finance: an overview of the issues and evidence. *Journal of Corporate Finance* **10**(2) 301–326. doi: 10.1016/S0929-1199(03)00059-2. Venture Capital, Initial Public Offerings, and Entrepreneurial Finance.

Desai, Akshar Prabhu, Tejasvi Ravi, Mohammad Luqman, Ganesh Mallya, Nithya Kota, Pranjul Yadav. 2024. Opportunities and challenges of generative-ai in finance. *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 4913–4920. doi: 10.1109/bigdata62323.2024.10825658.

Deutsche Bank. 2021. Most popular esg indices among corporate issuers and investors in the americas and europe in 2021 [graph]. `https://www.statista.com/statistics/1296728/most-used-esg-indices-in-the-americas-and-europe/`.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*. doi: 10.48550/arXiv.1810.04805.

Döll, Luciano Mathias, Micaela Ines Castillo Ulloa, Alexandre Zammar, Guilherme Francisco do Prado, Cassiano Moro Piekarski. 2022. Corporate venture capital and sustainability. *Journal of Open Innovation: Technology, Market, and Complexity* **8**(3) 132. doi: 10.3390/joitmc8030132.

El Hanchi, Samia, Lamia Kerzazi. 2020. Startup innovation capability from a dynamic capability-based view: A literature review and conceptual framework. *Journal of Small Business Strategy (archive only)* **30**(2) 72–92.

European Comission. 2022. Sustainability-related disclosure in the financial services sector. `https://finance.ec.europa.eu/sustainable-finance/disclosures/sustainability-related-disclosure-financial-services-sector_en`.

Fatemi, Ali M., Iraj J. Fooladi. 2013. Sustainable finance: A new paradigm. *Global Finance Journal* **24**(2) 101–113. doi: https://doi.org/10.1016/j.gfj.2013.07.006.

Fisch, Christian. 2019. Initial coin offerings (icos) to finance new ventures. *Journal of Business Venturing* **34**(1) 1–22. doi: 10.1016/j.jbusvent.2018.09.007.

Florysiak, David, Alexander Schandlbauer. 2019. The information content of ico white papers. *ERN: Foreign Exchange Models (Topic)* doi: 10.2139/ssrn.3265007.

Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5) 1189 – 1232. doi: 10.1214/aos/1013203451.

Gabbert, John, Nizar Tarhuni, Daniel Cook, Andrew Akers. 2022. Do vc industry specialists outperform? investigating performance differences between vc fund styles. URL `https://pitchbook.com/news/reports/q1-2022-pitchbook-analyst-note-do-vc-industry-specialists-outperform`.

Gao, Ziming, Yuan Gao, Yi Hu, Zhengyong Jiang, Jionglong Su. 2020. Application of deep q-network in portfolio management. *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*. 268–275. doi: 10.1109/ICBDA49040.2020.9101333.

Garkavenko, Mariia, Eric Gaussier, Hamid Mirisaee, Cédric Lagnier, Agnès Guerraz. 2022. Where do you want to invest? predicting startup funding from freely, publicly available web information. doi: 10.48550/arXiv.2204.06479.

Garkavenko, Mariia, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, Cédric Lagnier. 2021. Valuation of startups: A machine learning perspective. *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 176–189. doi: 10.1007/978-3-030-72113-8_12.

Géron, Aurélien. 2017. Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligentsystems. Paperback.

Gibson Brandon, R., P. Krueger, P. S. Schmidt. 2021. Esg rating disagreement and stock returns. *Financial Analysts Journal* **77**(4) 104–127. doi: 10.1080/0015198X.2021.1963186.

Giudici, Giancarlo, Saman Adhami. 2019. The impact of governance signals on ico fundraising success. *Journal of Industrial and Business Economics* **46**(2) 283–312. doi: 10.1007/s40812-019-00118-w.

Giudici, Giancarlo, Cristina Rossi-Lamastra. 2018. *Crowdfunding of SMEs and Startups: When Open Investing Follows Open Innovation*, chap. 12. 377–396. doi: 10.1142/9789813230972_0012.

Global Sustainable Investment Alliance. 2020. Global sustainable investment review 2020. `https://www.gsi-alliance.org/wp-content/uploads/2021/08/GSIR-20201.pdf`.

Global Sustainable Investment Alliance. 2021. Value of sustainable assets under management (aum) and total assets under management worldwide from 2016 to 2020. `https://www.statista.com/statistics/948492/value-sustainable-total-aum-worldwide/`.

Glücksman, Sarah. 2020. Entrepreneurial experiences from venture capital funding: exploring two-sided information asymmetry. *Venture Capital* **22** 331–354. doi: 10.1080/13691066.2020.1827502.

Gompers, Paul A. 1995. Optimal investment, monitoring, and the staging of venture capital. *The Journal of Finance* **50**(5) 1461–1489. doi: https://doi.org/10.1111/j.1540-6261.1995.tb05185.x.

Gompers, Paul A., Will Gornall, Steven N. Kaplan, Ilya A. Strebulaev. 2020. How do venture capitalists make decisions? *Journal of Financial Economics* **135**(1) 169–190. doi: 10.1016/j.jfineco.2019.06.011.

Gompers, Paul A., Anna Kovner, Josh Lerner. 2009. Specialization and success: Evidence from venture capital. *Journal of Economics & Management Strategy* **18**(3) 817–844. doi: 10.1111/j.1530-9134.2009.00230.x.

Guo, Tian, Nicolas Jamet, Valentin Betrix, Louis Alexandre Piquet, Emmanuel Hauptmann. 2020. Esg2risk: A deep learning framework from esg news to stock volatility prediction. doi: 10.48550/arXiv.2005.02527.

Gutierrez-Bustamante, Marcelo, Leonardo Espinosa-Leal. 2022. Natural language processing methods for scoring sustainability reportsmdash;a study of nordic listed companies. *Sustainability* **14**(15). doi: 10.3390/su14159165.

Hagenau, Michael, Michael Liebmann, Dirk Neumann. 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems* **55**(3) 685–697. doi: 10.1016/j.dss.2013.02.006.

Han, Bing, David Hirshleifer, Johan Walden. 2022. Social transmission bias and investor behavior. *Journal of Financial and Quantitative Analysis* **57**(1) 390–412. doi: 10.1017/S0022109021000077.

Hand, John R. M. 2001. The role of book income, web traffic, and supply and demand in the pricing of u.s. internet stocks. *European Finance Review* **5**(3) 295–317.

Hegeman, Puck D., Roger Sørheim. 2021. Why do they do it? corporate venture capital investments in cleantech startups. *Journal of Cleaner Production* **294** 126315. doi: 10.1016/j.jclepro.2021.126315.

Ho, Tina H. 2015. Social purpose corporations: the next targets for greenwashing practices and crowdfunding scams. *Seattle Journal for Social Justice* **13**(3) 14.

Hornuf, Lars, Theresa Kück, Armin Schwienbacher. 2022. Initial coin offerings, information disclosure, and fraud. *Small Business Economics* **58**. doi: 10.1007/s11187-021-00471-y.

Howell, Sabrina T, Marina Niessner, David Yermack. 2019. Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales. *The Review of Financial Studies* **33**(9) 3925–3974. doi: 10.1093/rfs/hhz131.

Hsu, David H. 2007. Experienced entrepreneurial founders, organizational capital, and venture capital funding. *Research Policy* **36**(5) 722–741. doi: https://doi.org/10.1016/j.respol.2007.02.022.

Hu, Yuh-Jong, Shang-Jen Lin. 2019. Deep reinforcement learning for optimizing finance portfolio management. *2019 Amity International Conference on Artificial Intelligence (AICAI)*. 14–20. doi: 10.1109/AICAI.2019.8701368.

Hutto, C., Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* **8**(1) 216–225. doi: 10.1609/icwsm.v8i1.14550.

ICObench. 2017. Rating methodology. `https://icobench.com/ratings`.

IFRS Foundation. 2022a. Materiality finder. `https://sasb.org/standards/materiality-finder`.

IFRS Foundation. 2022b. Sustainable development goals. `https://sasb.ifrs.org/standards/`.

Institute for Sustainable Investing. 2021. Sustainable Reality:2020 Update. `https://www.morganstanley.com/content/dam/msdotcom/en/assets/pdfs/3190436-20-09-15_Sustainable-Reality-2020-update_Final-Revised.pdf`.

Jiang, Zhengyao, Dixing Xu, Jinjun Liang. 2017. A deep reinforcement learning framework for the financial portfolio management problem. doi: 10.48550/arXiv.1706.10059.

Kaiser, Ulrich, Johan M. Kuhn. 2020. The value of publicly available, textual and non-textual information for startup performance prediction. *Journal of Business Venturing Insights* **14** e00179. doi: 10.1016/j.jbvi.2020.e00179.

Kane, Tim J. 2010. The importance of startups in job creation and job destruction. *Available at SSRN 1646934* .

Kanze, Dana, Laura Huang, Mark A Conley, E Tory Higgins. 2018. We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of management journal* **61**(2) 586–614.

Kaplan, Steven N., Josh Lerner. 2016. *Venture Capital Data: Opportunities and Challenges*. University of Chicago Press, 413–431.

Kaplan, Steven N., Per Strömberg. 2003. Financial Contracting Theory Meets the Real World: An Empirical Analysis of Venture Capital Contracts. *The Review of Economic Studies* **70**(2) 281–315. doi: 10.1111/1467-937X.00245.

Karpenko, Oksana A., Tatiana K. Blokhina, Lali V. Chebukhanova. 2021. The initial coin offering (ico) process: Regulation and risks. *Journal of Risk and Financial Management* **14**(12). doi: 10.3390/jrfm14120599.

Kessler, Alexander, Christian Korunka, Frank Hermann, Manfred Lueger. 2012. Predicting founding success and new venture survival: a longitudinal nascent entrepreneurship approach. *Journal of Enterprising Culture* **20**. doi: 10.1142/S0218495812500021.

Kharchenko, Stanislav, Mark Rebotunov, Daniel Afshar, Anton Matskevich. 2023. Understanding the multidimensional private equity space: Unsupervised learning for identifying relationships in the startup ecosystem. `https://terminal.twotensor.com/file/Private_Equity_Space.pdf`.

Kingma, Diederik P., Jimmy Ba. 2017. Adam: A method for stochastic optimization. doi: 10.48550/arXiv.1412.6980.

Köhn, Andreas. 2018. The determinants of startup valuation in the venture capital context: a systematic review and avenues for future research. *Management Review Quarterly* **68**(1) 3–36. doi: 10.1007/s11301-017-0131-5.

Korteweg, Arthur, Morten Sorensen. 2010. Risk and Return Characteristics of Venture Capital-Backed Entrepreneurial Companies. *The Review of Financial Studies* **23**(10) 3738–3772. doi: 10.1093/rfs/hhq050.

KPMG. 2022. Venture pulse q12022 global analysis of venture funding. `https://assets.kpmg/content/dam/kpmg/uk/pdf/2022/04/venture-pulse-q1-2022.pdf`.

Kraaijeveld, Olivier, Johannes De Smedt. 2020. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money* **65** 101188. doi: 10.1016/j.intfin.2020.101188.

Krappel, Tim, Alex Bogun, Damian Borth. 2021. Heterogeneous ensemble for ESG ratings prediction. *CoRR* **abs/2109.10085**. doi: 10.48550/arXiv.2109.10085.

Krishna, Amar, Ankit Agrawal, Alok Choudhary. 2016. Predicting the outcome of startups: Less failure, more success. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. 798–805. doi: 10.1109/ICDMW.2016.0118.

Kumar, Abhijeet. 2021. Finbert-embedding. `https://pypi.org/project/finbert-embedding`.

Kumar, Indu, Kiran Dogra, Chetna Utreja, Premlata Yadav. 2018. A comparative study of supervised machine learning algorithms for stock market trend prediction. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. 1003–1007. doi: 10.1109/ICICCT.2018.8473214.

Lahajnar, Sebastian, Alenka Rozanec. 2018. Initial coin offering (ico) evaluation model. *Investment Management and Financial Innovations* **15** 169–182. doi: 10.21511/imfi.15(4).2018.14.

Lawal, Zaharaddeen Karami, Hayati Yassin, Rufai Yusuf Zakari. 2020. Stock market prediction using supervised machine learning techniques: An overview. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. 1–6. doi: 10.1109/CSDE50874.2020.9411609.

Le, Quoc, Tomas Mikolov. 2014. Distributed representations of sentences and documents. Eric P. Xing, Tony Jebara, eds., *Proceedings of the 31st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 32. PMLR, Bejing, China, 1188–1196. doi: 10.48550/arXiv.1405.4053.

Lee, Jongsub, Tao Li, Donghwa Shin. 2021. The Wisdom of Crowds in FinTech:

Evidence from Initial Coin Offerings. *The Review of Corporate Finance Studies* **11**(1) 1–46. doi: 10.1093/rcfs/cfab014.

Lei, Yu, Wenjie Li. 2019. Interactive recommendation with user-specific deep reinforcement learning. *ACM Transactions on Knowledge Discovery from Data* **13**(6). doi: 10.1145/3359554.

Lerner, Josh, Ramana Nanda. 2020. Venture capital's role in financing innovation: What we know and how much we still need to learn. *Journal of Economic Perspectives* **34**(3) 237–61. doi: 10.1257/jep.34.3.237.

Li, Yinfei, Jiafeng Shou, Philip Treleaven, Jun Wang. 2022. Graph neural network for merger and acquisition prediction. *Proceedings of the Second ACM International Conference on AI in Finance*. ICAIF '21, Association for Computing Machinery, New York, NY, USA. doi: 10.1145/3490354.3494368.

Liebau, Daniel, Patrick Schueffel. 2019. Cryptocurrencies initial coin offerings: Are they scams? - an empirical study. *The Journal of the British Blockchain Association* **2** 1–7. doi: 10.31585/jbba-2-1-(5)2019.

Liebman, Elad, Peter Stone. 2014. DJ-MC: A reinforcement-learning agent for music playlist recommendation. *CoRR* **abs/1401.1880**. doi: 10.48550/arXiv.1401.1880.

Lin, Lin. 2022. Venture capital in the rise of sustainable investment. *European Business Organization Law Review* **23**(1) 187–216. doi: 10.1007/s40804-021-00238-8.

Lipusch, Nikolaus. 2018. Initial coin offerings a paradigm shift in funding disruptive innovation. *SSRN Electronic Journal* doi: 10.2139/SSRN.3148181.

Liu, Hannah M. 2019. Why do people invest in initial coin offerings (icos)? Bachelor Thesis. Joseph Wharton Scholars.

Liu, Xiao-Yang, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, Christina Dan Wang. 2022. Finrl: A deep reinforcement learning library for automated stock trading in qxuantitative finance. *SSRN Electronic Journal* doi: 10.2139/ssrn.3737859.

Lyandres, Evgeny, Berardino Palazzo, Daniel Rabetti. 2022. Initial Coin Offering (ICO) Success and Post-ICO Performance. *Management Science* **68**(12) 8658–8679. doi: 10.1287/mnsc.2022.4312.

Maats, F.H.E., Erasmus Universiteit. Faculteit Bedrijfskunde. 2008. *On the Consistency and Reliability of Venture Capital Databases*. Erasmus Universiteit.

Macmillan, Ian C., Robin Siegel, P.N.Subba Narasimha. 1985. Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business Venturing* **1**(1) 119–128. doi: 10.1016/0883-9026(85)90011-4.

Mansouri, Sasan, Paul P. Momtaz. 2022. Financing sustainable entrepreneurship: Esg measurement, valuation, and performance. *Journal of Business Venturing* **37**(6) 106258. doi: 10.1016/j.jbusvent.2022.106258.

Markowitz, Harry. 1952. Portfolio selection. *The Journal of Finance* **7**(1) 77–91.

Mayew, William J, Mohan Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *The Journal of Finance* **67**(1) 1–43.

McBride, Sarah. 2012. Facebook ipo has halo effect for venture capitalists. `https://www.reuters.com/article/markets/wealth/facebook-ipo-has-halo-effect-for-venture-capitalists-idUSBRE84C06J/`.

McKinsey. 2022. Mckinsey's private markets annual review. `https://www.mckinsey.com/industries/private-equity-and-principal-investors/our-insights/mckinseys-private-markets-annual-review`.

Miloud, Tarek, Arild Aspelund, Mathieu Cabrol. 2012. Startup valuation by venture capitalists: An empirical study. *Venture Capital: An International Journal of Entrepreneurial Finance* **14** 1–24. doi: 10.1080/13691066.2012.667907.

Minola, Tommaso, Marco Giorgino. 2008. Who's going to provide the funding for high tech start-ups? a model for the analysis of determinants with a fuzzy approach. *Ramp D Management* doi: 10.1111/j.1467-9310.2008.00518.x.

Mittermayer, M.-A. 2004. Forecasting intraday stock price trends with text mining techniques. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. 10. doi: 10.1109/HICSS.2004.1265201.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin A. Riedmiller. 2013. Playing atari with deep reinforcement learning. *CoRR* **abs/1312.5602**. doi: 10.48550/arXiv.1312.5602.

Momtaz, Paul P. 2020. Initial coin offerings. *Plos one* **15**(5) e0233018. doi: 10.1371/journal.pone.0233018.

Monika Dhochak, Sudesh Pahal, Prince Doliya. 2024. Predicting the startup valuation: A deep learning approach. *Venture Capital* **26**(1) 75–99. doi: 10.1080/13691066.2022.2161968.

Montani, Damiano, Daniele Gervasio, Andrea Pulcini. 2020. Startup company valuation: The state of art and future trends. *International Business Research* **13** 31. doi: 10.5539/ibr.v13n9p31.

Moody, J., M. Saffell. 2001. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks* **12**(4) 875–889. doi: 10.1109/72.935097.

Mousavi, Seyed Sajad, Michael Schukat, Enda Howley. 2018. Deep reinforcement learning: An overview. Yaxin Bi, Supriya Kapoor, Rahul Bhatia, eds., *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*. Springer International Publishing, Cham, 426–440. doi: 10.1007/978-3-319-56991-8_32.

Mukherjee, Mukut. 2020. Esg-bert: Nlp meets sustainable investing. `https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b`.

Myers, Stewart C. 1977. Determinants of corporate borrowing. *Journal of Financial Economics* **5**(2) 147–175. doi: 10.1016/0304-405X(77)90015-0.

Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven D. Brown. 2004. An introduction to decision tree modeling. *Journal of Chemometrics* **18**(6) 275–285. doi: https://doi.org/10.1002/cem.873.

Nanda, Ramana, Matthew Rhodes-Kropf. 2013. Investment cycles and startup innovation. *Journal of Financial Economics* **110**(2) 403–418. doi: 10.1016/j.jfineco.2013.07.001.

Narayanan, Arvind, Joseph Bonneau, Edward Felten, Andrew Miller, Steven Goldfeder. 2016. *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press.

Nielsen, Finn Årup. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. doi: 10.48550/arXiv.1103.2903.

Norton, Edgar, Bernard H. Tenenbaum. 1993. Specialization versus diversification as a venture capital investment strategy. *Journal of Business Venturing* **8**(5) 431–442. doi: https://doi.org/10.1016/0883-9026(93)90023-X.

Nugent, Tim, Nicole Stelea, Jochen L. Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 157–169. doi: 10.1007/978-3-030-86967-0_12.

OECD. 2024. Measuring job creation by start-ups and young firms. `https://www.oecd.org/en/about/projects/measuring-job-creation-by-start-ups-and-young-firms.html`.

Ofir, Moran, Ido Sadeh. 2020. Ico vs. ipo: Empirical findings, information asymmetry, and the appropriate regulatory framework. *Vand. J. Transnat'l L.* **53** 525. doi: 10.2139/ssrn.3338067.

Pagolu, Venkata Sasank, Kamal Nayan Reddy, Ganapati Panda, Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. *2016*

International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). 1345–1350. doi: 10.1109/SCOPES.2016.7955659.

Palma, Gabriel Rodrigues, Mariusz Skoczeń, Phil Maguire. 2024. Combining supervised and unsupervised learning methods to predict financial market movements. URL `https://arxiv.org/abs/2409.03762`.

Park, Hyungjun, Min Kyu Sim, Dong Gu Choi. 2020. An intelligent financial portfolio trading strategy using deep q-learning. *Expert Systems with Applications* **158** 113573. doi: 10.1016/j.eswa.2020.113573.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Pavlidis, N. G., V. P. Plagianakos, D. K. Tasoulis, M. N. Vrahatis. 2006. Financial forecasting through unsupervised clustering and neural networks. *Operational Research* **6**(2) 103–127. doi: 10.1007/BF02941227.

Payne, Bill. 2011. Valuations 101: The dave berkus method. `https://gust.com/blog/valuations-101-the-dave-berkus-method/`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** 2825–2830. doi: 10.48550/arXiv.1201.0490.

Pilkington, Marc. 2018. The emerging ico landscape - some financial and regulatory standpoints. *SSRN Electronic Journal* doi: 10.2139/ssrn.3120307.

Pitchbook. 2022. Meta company profile. Retrieved from Pitchbook database.

Pitchbook. 2024. Gocardless company profile. Retrieved from Pitchbook database.

Powell, Nicole, Simon Y. Foo, Mark Weatherspoon. 2008. Supervised and unsupervised methods for stock trend forecasting. *2008 40th Southeastern Symposium on System Theory (SSST)*. 203–205. doi: 10.1109/SSST.2008.4480220.

Preqin. n.d. Esg solutions | preqin. `https://www.preqin.com/esg/esg-solutions`.

Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, Igor Mozetič. 2015. The effects of twitter sentiment on stock price returns. *PloS one* **10**(9) e0138441. doi: 10.1371/journal.pone.0138441.

Refinitiv. 2022. Environmental, social and governance scores from refinitiv. `https://www.lseg.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf`.

Rehurek, Radim, Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* **3**(2).

Reimers, Nils, Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. doi: 10.48550/arXiv.1908.10084.

Ressin, Marat. 2022. Start-ups as drivers of economic growth. *Research in Economics* **76**(4) 345–354. doi: 10.1016/j.rie.2022.08.003.

Retterath, Andre, Reiner Braun. 2020. Benchmarking venture capital databases. *SSRN Electronic Journal* doi: 10.2139/ssrn.3706108. Working Paper.

Reverte, C., M. D. Hernandez, A. Rojo-Ramírez. 2016. The profile of venture capital investments: The european context. *International Journal of Business and Globalisation* **17** 83–110. doi: 10.1504/IJBG.2016.077568.

Rooney, Kate. 2018. Ethereum falls on report that the second-biggest cryptocurrency is under regulatory scrutiny. `https://www.cnbc.com/2018/05/01/`

ethereum-falls-on-report-second-biggest-cryptocurrency-is-under-regulatory-s html.

Ross, Greg, Sanjiv Das, Daniel Sciro, Hussain Raza. 2021. Capitalvx: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science* **7** 94–114. doi: 10.1016/j.jfds.2021.04.001.

Roure, Miriam. 2024. SFDR, a primer for venture capital investors fundraising in the EU. https://www.vcstack.io/blog/sfdr-a-primer-for-venture-capital.

Ruberg, Nicolaas, Paolo Torroni, Vanessa Almeida. 2021. Bert goes sustainable: an nlp approach to esg financing. Master's thesis, Alma Mater Studiorum Università di Bologna.

Salesforce. 2021. Salesforce completes acquisition of slack. https://www.salesforce.com/uk/news/press-releases/2021/07/21/salesforce-slack-deal-close/.

Sander, Priit, Margus Kõomägi. 2007. Valuation of private companies by estonian private equity and venture capitalists. *Baltic Journal of Management* **2** 6–19. doi: 10.1108/17465260710720219.

Sapkota, Niranjan, Klaus Grobys, Josephine Dufitinema. 2020. How much are we willing to lose in cyberspace? on the tail risk of scam in the market for initial coin offerings. *SSRN Electronic Journal* doi: 10.2139/ssrn.3732747.

Sazzed, Salim, Sampath Jayarathna. 2021. Ssentia: A self-supervised sentiment analyzer for classification from unlabeled data. *Machine Learning with Applications* **4** 100026. doi: 10.1016/j.mlwa.2021.100026.

Schilling, Melissa. 2002. Technology success and failure in winner-take-all markets: The impact of learning orientation, timing, and network externalities. *Academy of Management Journal* **45** 387–398. doi: 10.2307/3069353.

Schmidt, Alexander. 2019. Sustainable news–a sentiment analysis of the effect of esg information on stock prices. *SSRN Electronic Journal* doi: 10.2139/ssrn.3809657.

Schumaker, Robert P., Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.* **27**(2). doi: 10.1145/1462198.1462204.

Serafeim, George, Aaron Yoon. 2022. Which corporate esg news does the market react to? *Financial Analysts Journal* **78**(1) 59–78. doi: 10.1080/0015198X.2021.1973879.

Serafeim, George, Aaron Yoon. 2023. Stock price reactions to esg news: the role of esg ratings and disagreement. *Review of Accounting Studies* **28**(3) 1500–1530. doi: 10.1007/s11142-022-09675-3.

Shanaev, Savva, Binam Ghimire. 2022. When esg meets aaa: The effect of esg rating changes on stock returns. *Finance Research Letters* **46** 102302. doi: 10.1016/j.frl.2021.102302.

Sharchilev, Boris, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, Maarten de Rijke. 2018. Web-based startup success prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM ’18, Association for Computing Machinery, New York, NY, USA, 2283–2291. doi: 10.1145/3269206.3272011.

Shevlin, Terry. 1996. The value-relevance of nonfinancial information: A discussion. *Journal of Accounting and Economics* **22**(1) 31–42. doi: 10.1016/S0165-4101(96)00441-7.

Silva Júnior, Claudio Roberto, Julio Cezar Mairesse Siluk, Alvaro Neuenfeldt Júnior, Carmen Brum Rosa, Cláudia de Freitas Michelin. 2022. Overview of the factors that influence the competitiveness of startups: a systematized literature review. *Gestão & Produção* **29** e13921.

Smith, Richard, Robert Pedace, Vijay Sathe. 2011. Vc fund financial performance: The relative importance of ipo and ma exits and exercise of abandonment options. *Financial Management* **40**(4) 1029–1065. doi: 10.1111/j.1755-053X.2011.01170.x.

Sohangir, Sahar, Nicholas Petty, Dingding Wang. 2018. Financial sentiment lexicon analysis. *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. 286–289. doi: 10.1109/ICSC.2018.00052.

Śpiewanowski, Piotr, Oleksandr Talavera, Linh Vi. 2022. Applications of web scraping in economics and finance. *Oxford Research Encyclopedia of Economics and Finance*.

Spooner, Thomas, John Fearnley, Rahul Savani, Andreas Koukorinis. 2018. Market making via reinforcement learning. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 434–442.

Stanley, Maxwell. 2019. The application of behavioural heuristicsto initial coin offerings valuation and investment. *The Journal of the British Blockchain Association* **2** 1–7. doi: 10.31585/JBBA-2-1-(7)2019.

Statista Research Department. 2024. Number of jobs created by start-up businesses that were less than one year old in the united states from 1994 to 2023. `https://www.statista.com/statistics/235515/jobs-created-by-start-ups-in-the-us/`.

Summers, Ed, Igor Brigadir, Sam Hames, Hugo van Kemenade, Peter Binkley, tinafigueroa, Nick Ruest, Walmir, Dan Chudnov, recrm, celeste, Hause Lin, Andy Chosak, R. Miles McCain, Ian Milligan, Andreas Segerberg, Daniyal Shahrokhian, Melanie Walsh, Leonard Lausen, Nicholas Woodward, Felix Victor Münch, eggplants, Ashwin Ramaswami, Darío Hereñú, Dmitrijs Milajevs, Frederik Elwert, Kalle Westerling, rongpenl, Stefano Costa, Shawn. 2014. twarc. `https://github.com/DocNow/twarc/tree/v2.10.4`.

Sutton, Richard S, Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Taghipour, Nima, Ahmad Kardan, Saeed Shiry Ghidary. 2007. Usage-based web recommendations: A reinforcement learning approach. *Proceedings of the 2007 ACM Conference on Recommender Systems*. RecSys '07, Association for Computing Machinery, New York, NY, USA, 113–120. doi: 10.1145/1297231.1297250.

Tam, Pui-Wing, Shayndi Raice. 2012. A $9 billion jackpot for facebook investor. `https://www.wsj.com/articles/SB10001424052970204573704577186813800414598`.

Taçoğlu, Caner. n.d. Whitelist. `https://academy.binance.com/en/glossary/whitelist`.

Thomas, Meyer, Mathonet Pierre-Yves. 2005. *Beyond the J Curve : Managing a Portfolio of Venture Capital and Private Equity Funds.*. [Wiley Finance], Wiley.

Tran, Hanah, Patrick J Murphy. 2023. Generative artificial intelligence and entrepreneurial performance. *Journal of Small Business and Enterprise Development* **30**(5) 853–856. doi: 10.1108/JSBED-09-2023-508.

Trevor Rogers, F. 2020. Patent text similarity and cross-cultural venture-backed innovation. *Journal of Behavioral and Experimental Finance* **26** 100319. doi: 10.1016/j.jbef.2020.100319.

United Nations. 2015. Sustainable development goals. `https://www.undp.org/sustainable-development-goals`.

U.S. Bureau of Labor Statistics. 2022. Survival of private sector establishments by opening year. `https://www.bls.gov/bdm/us_age_naics_00_table7.txt`.

Venuti, Keenan. 2021. Predicting mergers and acquisitions using graph-based deep learning. *CoRR* **abs/2104.01757**. doi: 10.48550/arXiv.2104.01757.

Vismara, Silvio. 2019. Sustainability in equity crowdfunding. *Technological Forecasting and Social Change* **141** 98–106. doi: 10.1016/j.techfore.2018.07.014.

Wang, Susheng, Hailan Zhou. 2004. Staged financing in venture capital: moral hazard and risks. *Journal of Corporate Finance* **10**(1) 131–155. doi: 10.1016/S0929-1199(02)00045-7.

Watkins, Christopher J. C. H., Peter Dayan. 1992. Q-learning. *Machine Learning* **8**(3) 279–292. doi: 10.1007/BF00992698.

Wei Shi, Savannah. 2018. Crowdfunding: Designing an effective reward structure. *International Journal of Market Research* **60**(3) 288–303. doi: 10.1177/1470785317744113.

Wiek, Hilary, Anikka Villegas, TJ Mei, Sarah Schwab, Julia Midkiff. 2023. 2023 sustainable investment survey. `https://pitchbook.com/news/reports/2023-sustainable-investment-survey`.

Williams, John Burr. 1938. *The Theory of Investment Value*. Harvard University Press.

Wilson, Theresa, Janyce Wiebe, Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05, Association for Computational Linguistics, USA, 347–354. doi: 10.3115/1220575.1220619.

Xiong, Hong, Ying Fan. 2021. How to better identify venture capital network communities: Exploration of a semi-supervised community detection method. *Journal of Social Computing* **2**(1) 27–42. doi: 10.23919/JSC.2020.0008.

Yang, Zhenjia, Daniel Broby. 2020. Sustainable finance : Ai applications in satellite imagery and data. Discussion paper, Glasgow. URL `https://strathprints.`

`strath.ac.uk/73828/`. Financial technology paper, published as part of the Centre for Financial Regulation and Innovation, Strathclyde Business School, University of Strathclyde.

Yu, Ellen Peiyi, Bac Van Luu, Catherine Huirong Chen. 2020. Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance* **52** 101192. doi: 10.1016/j.ribaf.2020.101192.

Zhang, Ruling, Zengrui Tian, Killian J. McCarthy, Xiao Wang, Kun Zhang. 2023. Application of machine learning techniques to predict entrepreneurial firm valuation. *Journal of Forecasting* **42**(2) 402–417. doi: 10.1002/for.2912.

Zhang, Xubo. 2012. Venture capital investment selection decision-making base on fuzzy theory. *Physics Procedia* **25** 1369–1375. doi: 10.1016/j.phpro.2012.03.248. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.