Universität St.Gallen

Investments in Emerging Technology:

Education.ai – Exploring the Influence of Artificial Intelligence Adoption on Investor Behavior in the EdTech Sector

Thesis for the attainment of the degree Master of Arts

Supervised by	Dr. Naomi Haefner & Prof. Dr. Oliver Gassmann Chair for Innovation Management Institute of Technology Management University of St.Gallen
Composed by	Mario Sebastian Muehlematter Master of Arts in Unternehmensführung (MUG-HSG) 17-603-440
Submission	St.Gallen / 20. Nov. 2023

ABSTRACT

A dataset containing information about 14’046 businesses operating in the Education Technology (EdTech) market was investigated regarding the adoption of Artificial Intelligence (AI). The ventures contained in the dataset could be assigned to 23 topics, however, none is labeled “AI”. The keywords representing said topics mostly reflect the primary visions and customer segments. The analysis suggests that AI is only a tool to realize these visions, therefore about the operational dimension of the business model. The idea of AI-first EdTech companies does not appear in the self-description of ventures listed on CrunchBase. The logical conclusion of this first analysis is that AI is not the “hot” topic in EdTech, as it is not a topic at all. A binary classification was added to the dataset in the second analytical step. It confirmed that **Artificial Intelligence (AI) is not the current “hot” topic in EdTech,** as the distribution of AI companies regarding size, total funding, and age appears random. As the data revealed no trend, analyzing investor behavior regarding a trend-based bias linked to AI was made impossible. However, the temporal analysis using Kernel Density Estimates (KDE) on the topics resulting from the first step revealed additional insights linked to the wider field of emerging technologies: In the past Virtual Reality (VR) was a “hot” topic, explicitly referring to a specific technology, in the EdTech market. **The association with this “hot” topic influenced investor behavior negatively**, the ventures founded during the hype phase with growing interest in VR received less funding on average than both VR business across the entire dataset and the wider EdTech peer group founded within the same timeframe. This finding partly validates the concept of “smart money”, as professional investors do not overinvest based on a trend-driven bias.

Contents

Figures

Abbreviations

AI	Artificial Intelligence
BoW	Bag-of-Words
CB	CrunchBase; crunchbase.com
chpt.	Chapter
DCF	Discounted Cash Flow
EdTech	Education Technology
Fig.	Figure
FTE	Full-Time Employee
KDE	Kernel Density Estimate
LLM	Large Language Model
M&A	Mergers & Acquisitions
ML	Machine Learning
NLP	Natural Language Processing
RFC	Random Forest Classifier
USD	United States Dollar
VC	Venture Capital
VR	Virtual Reality
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Topic Modeling

1 INTRODUCTION

"Overall, a mere association with the Internet seems enough to provide a firm with a large and permanent value increase."

(Cooper et al., 2001, p. 2386).

This surprising insight from researching the dot-com bubble of the early 2000s, seemingly running against conventional business logic, is the inspiration for this thesis. The underlying idea can be more broadly framed as exploring the relationship between topics and business valuation. The basic idea is to identify "hot" or "cold" topics. In scientific literature topics with growing interest, thus "hot" topics, are delimited by an increasing number of publications, whilst "cold" topics are the inverse. (Griffiths & Steyvers, 2004). This definition intended for scientific publications can be easily adapted for start-ups by replacing the number of articles mentioning a certain topic with the number of newly founded firms (self-)identifying with a topic employing their description. In the contemporary technological innovation landscape, perhaps no field garners as much attention and fascination as Artificial Intelligence (AI). The rise of AI technologies has been nothing short of meteoric, with its influence extending far beyond computer science to profoundly impact various aspects of our daily lives. Illustrative of this rise to prominence is the increased search interest for the term (Google, 2023). The Gartner EmTech Report is an annual publication from a leading technology consultancy, thus underpinning the relevance of the technology from a practitioner's view. It lists multiple forms of AI in the adoption cycle. At the same time, the current version of the publication is also a warning of inflated expectations as most topics relating to AI are at the "Peak of Inflated Expectations". (Gartner, 2023).

OpenAI, a leading player in the AI industry, has shattered records in terms of adoption speed, reaching 1 million users within a week (UBS, 2023). The pace at which OpenAI's technologies have been embraced is indicative of the fervent curiosity and eagerness of both individuals and organizations to harness the capabilities of AI for solving complex problems and driving innovation. This unprecedented adoption rate reflects not only the transformative potential of AI but also the compelling need for its integration across diverse sectors of society, including education.

In contrast to the beforementioned study tracking name changes referring to the internet in public markets (Cooper et al., 2001), this thesis narrows the scope to derive deep insights into Education Technology (EdTech) companies. Revisiting a concept already established more than 20 years ago is merited by the shift in hotly discussed technologies, today AI is what the internet once represented. As already touched upon AI is a rapidly growing field that has seen an influx of investment and innovation over the past decade. This deeply rooted analysis of a private market also incorporates emerging and less-known players in the domain. Leveraging machine learning, I also extend the body of analysis and aim to integrate the company description rather than solely their name to determine association with a topic. This shift in scope is made possible by two distinct advances in computing. Firstly, data availability has increased significantly, enabling an analysis of private markets. All major providers of private market data for Venture Capital (VC) have been founded after 2001. Their rise was significantly aided by the increased interest in private markets. (Retterath & Braun, 2020, p. 2). This thesis relies on CrunchBase as a data provider, as is considered reliable and accessible for researchers. Secondly, what must have seemed like an impossible amount of data to process in the early 2000’s, is well within the possibilities of modern widely available computers and cloud infrastructure. The rise of cloud computing has revolutionized how businesses and individuals access computing resources. At the heart of this development is Moore's Law, which approximates that the number of transistors on a microchip doubles every two years (Moore, 1998). This has allowed for the development of ever more powerful and efficient computing systems, which have made cloud computing possible. In turn, cloud computing has allowed for the creation of new applications and services that would have been impossible just a few years ago, making it an essential part of the modern technology landscape.

The described rise of new topics and tools culminates in the following research questions:

RQ1: Is Artificial Intelligence (AI) the current "hot" topic in EdTech?

RQ2: Does association with a "hot" topic influence investor behavior?

This thesis is structured as follows. Firstly, existing literature is condensed, and important terminology is explained. This is followed by the methodology part highlighting topic modeling (TM) and classification methods. As well as explaining data collection and treatment. The results part of this thesis reveals the key insights regarding investor behavior in the EdTech market, which are subsequentially discussed.

2 THEORETICAL BACKGROUND

This chapter contains two major subchapters, each treating a fundamental theoretical field linked to this thesis. In order these fields are Education Technology (EdTech) and Investor Behavior. The first part describes the objects to which investors allocate funds, and the second part their approach to valuing said object.

2.1 EDUCATION TECHNOLOGY

EdTech, short for Educational Technology, refers to the use of technology and digital tools to enhance and support teaching and learning processes. EdTech, as understood in this thesis, encompasses a wide range of technologies and digital resources, including software, hardware, online platforms, and applications designed for educational purposes.

2.1.1 THE EDTECH MARKET

Education, a highly traditional industry, is currently experiencing a profound disruption driven by the forces of digitalization. This trend has been accelerated by the COVID-19 pandemic. (Gallagher & Palmer, 2020). On a global scale, almost immediately students had to switch from on-site to digital ways of learning, subverting long-establish, traditional ways of teaching. EdTech has remained relevant ever since. (Lyons, 2017). The EdTech market has been experiencing strong growth with increased investments from venture capitalists as well as government initiatives aimed at promoting EdTech adoption in schools around the world. Driven by the underlying VC boom and the pandemic pushing education online, private investment in EdTech grew from 16.1 billion USD in 2020 to 20.8 billion USD in 2021. However, in 2022 the market started to cool a little with investments falling to 10. 6 billion USD. The predictions for 2023 suggest total investments in the pre-pandemic range between 7 and 8.5 billion USD. (Whitford, 2023).

2.1.2 EMERGING BUSINESS MODELS

Whilst the market definition above clarifies the wider context a business operates in, it is not specific enough to distinguish between businesses. In the field of business strategy and innovation, a business model serves as a fundamental framework for answering critical questions that underpin the essence of a business. Gassmann, Frankenberger, and Csik's (2014) framework, which organizes a business model into four dimensions, forms a cornerstone of this thesis. Their four dimensions encompass the "WHO," "WHAT," "HOW," and "WHY" aspects of a business model, collectively providing a comprehensive view of the business's structure and functionality. The "WHO" dimension pertains to the identification of the target customer group, a fundamental consideration for any business model. The "WHAT" dimension revolves around the value proposition, delineating what the business offers and what the customer values. To execute these value propositions, the "HOW" dimension emphasizes the need for mastery of various processes, activities, resources, and capabilities. Lastly, the "WHY" dimension delves into the financial viability of the business model, encompassing aspects like cost structures and revenue mechanisms. This holistic framework serves as a vital tool in dissecting, understanding, and innovating business models, guiding the strategic decisions and value creation efforts of organizations. (Gassmann et al., 2014, pp. 8–12, 2016, pp. 19–22).

An important sub-category of businesses popularizing emerging technologies is start-ups. Whilst there are numerous definitions of start-ups to be found, in this thesis a relatively simple definition is chosen: All firms younger than 2 Years that did not form through M&A are considered a start-up. Besides its simplicity, the main advantage of this version is the overlap with the OECD (Criscuolo et al., 2014, p. 19), thus enabling comparison with established economic facts. Furthermore, this definition facilitates a simple temporal subsampling when analyzing the dataset later in this thesis.

2.1.3 ADOPTION OF ARTIFICIAL INTELLIGENCE

AI in the context of this thesis is treated in depth for two reasons. It appears as the most probable “hot” topic to a casual observer, and it underpins the NLP analysis from a technological standpoint. All information about the use of AI as a tool in this thesis can be found in Chapter 3.2.

The term AI is a colloquial umbrella term masking a wealth of different technologies and concepts. Given the very complex nature of the underlying technology, I introduce the fundamental pillars of AI and related concepts in the following chapter. In this chapter first, a short overview of the philosophical concepts underpinning AI is given, followed by an overview of the technological evolution and lastly, an assessment of the impact on the EdTech sector is provided.

The creation of (artificial) humans is by no means a new topic of interest, virtually all civilizations have their legend or religious origin story of humans. Be it the creation of man on the 6th day in the Abrahamitic religions or Prometheus, who shaped man out of mud, and was aided by Athena, who breathed life into the clay figure.

The Turing Test, originally named “the imitation game”, can be considered the modern focal point of this line of thought asking: What constitutes a human being? Named after his creator, who also created the Enigma decryption machine, which is widely accepted as one of the crucial predecessors to modern computers, the Turing Test remains relevant to the contemporary discussion of computing. In simple terms, the test declares artificial intelligence, in the terminology of the original, thinking machines, as equal to human intelligence if you can converse with it without noticing the otherness of the machine. (Oppy & Dowe, 2021; Turing, 1950).

Modern computer technology is on the verge of crossing the boundaries set forth by Turing. In a comprehensive gamified online study[1] modeled after Turing's thought experiment involving 1.5 million users participating in anonymous two-minute chat sessions, a fascinating pattern emerged. Participants would either chat with a fellow participant or with LLM-enabled chatbots, they could correctly identify the identity of their chat partners as either human or non-human in only 68% of the games. The results not only underscore the inherent difficulty in gauging the authenticity of online interactions but also reveal the potential limitations of human intuition when confronted with increasingly sophisticated AI technologies. The lines between human and artificial intelligence seem increasingly blurry. (Jannai et al., 2023).

However, there are challenges to the validity of the Turing test as a delimiter between human and artificial intelligence. Chief among them is the Chinese Room Argument, standing as one of the most influential and contentious thought experiments in the philosophy of mind and artificial intelligence. The thought experiment was first published in a 1980 article titled "minds, brains, and programs" by American philosopher John Searle.

In essence, Searle challenges the notion that a computer program can possess genuine understanding or consciousness. He presents a scenario in which a person, who does not understand Chinese, is situated in a room with a set of instructions for manipulating Chinese symbols. Outsiders pass questions written in Chinese into the room, and the person, following the instructions, generates responses that appear as if they understand Chinese.

Searle's argument hinges on the assertion that merely manipulating symbols according to a set of rules does not constitute genuine understanding. He contends that, despite appearances, the person in the room does not comprehend the Chinese

[1] https://www.humanornot.ai/

language any more than a computer running a program does. This thought experiment underscores Searle's belief that syntax, thus the manipulation of symbols, alone is insufficient for semantics. (Searle, 1980).

Critics and proponents alike engage in ongoing debate over the Chinese Room Argument's validity and implications for artificial intelligence and the philosophy of consciousness. While it does not provide a definitive conclusion, Searle's argument has undeniably sparked significant discourse and remains a touchstone in these fields, challenging us to grapple with the fundamental question of what it means to truly understand and think. (Cole, 2023).

The opinions in the academic discourse concerning consciousness range, in short, from positing it as the driving force explaining even the very laws of nature, the so-called Panpsychism (Goff, 2020) to declaring it a mere illusion, suggesting that the human brain is incapable of understanding its inner workings and thus infers that some “magical” ingredient, in this case, consciousness, must exist although it does not (Dennett, 2016).

For Chomsky et al. (2023) one of the decisive factors in considering AI sub-par to human intelligence is that although it appears capable of sophisticated thought and language, it shows the moral indifference born of unintelligence. The memory-based and statistical nature of AI means that it lacks another hallmark of intelligence too: The ability to create ideas and theories with universal reach. Furthermore, they argue that contemporary AI systems are far from rivaling humans in efficiency, as human minds can easily create explanations without the need to infer brute correlations among data points.

Others find a convergence of Brains and Algorithms, thus implying that humans and machines are not dissimilar in their reasoning. (Caucheteux & King, 2022). However, this convergence is not a given for all systems, as other researchers describe the comparison resulting in stark contrast: “the non-human-like nature of the knowledge learned by the network." became obvious (Mitchell & Bowers, 2020, p. 5147).

The term Artificial Intelligence was coined in a research proposal by McCarthy, Minsky, Rochester, and Shannon (1955). Minsky would later become one of the leading thinkers in the field. Early AI research included the perceptron: a rudimentary image classifier hailed in news articles as an artificial brain (Mason et al., 1958; Rosenblatt, 1958). The idea of imitating the human brain is pervasive in the history of AI. The fundamental idea to imitate the human brain and create sub-units of decision-making would later culminate in the development of neural networks. The early enthusiasm for the topic however faded and gave way to the so-called "AI winter", a period during which progress stalled. Even decades after the initial interest in the topic, research, such as early advances in self-driving vehicles, was severely limited by the computational power of its time (Pomerleau, 1988).

The advances of modern AI, the so-called "AI spring", start with the advent of big data. Independently, but importantly contemporary to the rise of Big Data, the idea of borrowing structurally from biology, especially the human brain, was reinvigorated, and Neural Networks emerged.

An early driver of the new age of AI was the ImageNet challenge. The challenge makes use of big data, providing a big collection of labeled images to train and test image classification algorithms. The basic concept was to create an algorithm that could analyze a random picture and correctly classify whatever it was depicting (Russakovsky, Deng et al., 2015).

The ImageNet challenge culminated in the emergence of an algorithm named AlexNet as the ultimate victor. AlexNet's triumph in the competition marked a significant turning point in the domain of image classification and recognition. This pioneering convolutional neural network, designed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, demonstrated an unprecedented breakthrough in the ability to recognize objects and images within vast datasets. Its success was a result of several innovative elements, chief among them the adoption of GPU-based parallel computing to expedite model training. AlexNet's remarkable accuracy and efficiency in processing vast image databases not only secured its victory in the ImageNet challenge but also set the stage for the subsequent deep learning revolution. (Krizhevsky et al., 2017).

The impact of AlexNet reverberated across various applications, including image recognition, natural language processing, and autonomous systems, reshaping the landscape of artificial intelligence and establishing deep learning as a dominant paradigm in the AI community.

The introduction of the transformer architecture marked another seminal breakthrough in the field of AI. The transformer model revolutionized the treatment of sequential data, particularly in tasks involving language understanding and generation. This innovative architecture introduced a mechanism based on self-attention. Self-attention enabled the model to weigh the importance of different words in a sentence, allowing it to capture complex relationships and long-range dependencies in the data. (Vaswani et al., 2017).

The introduction of the transformer architecture not only significantly improved the performance of various NLP tasks but also paved the way for the development of large pre-trained language models like BERT, GPT, and their derivatives, which have become the backbone of state-of-the-art models with enormous sizes in terms of parameters (Sevilla et al., 2022; Thompson, 2021).

AI-driven applications have become an integral part of our digital landscape, transforming the way humans interact with technology in virtually all industries (Chui et al., 2018). ChatGPT's rapid rise is emblematic of the increasing role AI plays in our daily lives, as users embrace AI for a multitude of applications, from virtual assistants to content generation. (UBS, 2023).

The advent of AI has undeniably ushered in a multitude of opportunities, but companies face diverse hurdles in realizing efficiency and efficacy gains (Haefner et al., 2023; Raisch & Krakowski, 2021). Furthermore, AI is also riddled with controversies and social issues. A prominent concern within the realm of AI is the manifestation of bias in AI systems. In one striking example, Google misclassified pictures of black people as gorillas (Barr, 2015).

Bias can creep into AI systems through various means, including biased training data, algorithmic biases, and biased decision-making processes. These biases can lead to discriminatory outcomes, unequal access, and perpetuation of societal inequalities. (Cowgill, 2019).

Furthermore, the proliferation of AI carries wide-ranging implications for the workforce. The deployment of AI technologies has the potential to transform industries and job landscapes, posing challenges in terms of workforce adaptability and implications for overall employment. (Berg et al., 2018; Brynjolfsson & McAfee, 2017; Brynjolfsson & Mitchell, 2017).

Besides social downsides, there is also the fear of dual use of AI, as it is finding application both as a military and civilian technology. In one example a team of researchers normally researching molecules for pharmaceutical use managed to generate the structure of 40,000 poisons, amongst them the chemical agent VX, which finds use as a chemical weapon. This sinister aim was achieved by the relatively simple reversal of one variable in their otherwise civilian algorithm serving medical research. (Urbina et al., 2022).

More direct military applications include target identification and designation, depending on military doctrine even probability-based life/death decisions are gaining prominence (Michel, 2023).

Not only are the future implications of the technology unclear but also the evolution of models itself is unclear. The evolution of AI models, particularly exemplified by

ChatGPT, has introduced a level of unpredictability in their behavior. Small, seemingly inconsequential modifications to input can yield unexpected outcomes, showcasing the intricacies of model behavior and the potential for inconsistencies. ChatGPT has displayed shifts in its outputs over time, emphasizing the challenge of ensuring reliable and consistent performance. (Chen et al., 2023).

Lastly the exponential increase in transistor density, colloquially called Moore's Law, an underlying driver of the current "AI spring" is potentially reaching its end as the physical components of computers are quickly nearing the atomic limit, the theoretical minimum size of one atom to conduct electric current (Waldrop, 2016). This implies that without advances in quantum computing, further advances in AI based on increasing model size will come to an end as the physical computer parts are no longer enabling this strategy for model improvement.

The range of technologies that constitute AI includes rule-based systems, classic machine learning, and representation learning (Goodfellow et al., 2016). Other definitions and terminology exist, regularly classic machine learning is referred to as supervised (machine) learning, and representation learning is referred to as unsupervised (machine) learning.

The working definition of AI for this thesis is wide. This means that all forms of "algorithmic intelligence" mentioned above are considered.

This has primarily a practical reason: the differentiation into more fine-grained categories requires a huge amount of data, with precise tags. This kind of dataset is not easily available. The pre-existing dataset that focuses on identifying AI in a management science context uses a high-level definition (Miric et al., 2023).

The education sector, a historically conventional industry, is undergoing a substantial transformation, largely catalyzed by the pervasive influence of digitalization. The COVID-19 pandemic accelerated this shift, as students worldwide swiftly transitioned from traditional, in-person learning to digital modalities. This abrupt change challenged established teaching methodologies and highlighted the significance of EdTech. The enduring relevance of digital solutions in education signifies a lasting impact on how education is delivered and accessed and the continued evolution of learning paradigms in response to changing global dynamics. signifying a lasting impact on how education is delivered and accessed. (Gallagher & Palmer, 2020)

At the heart of this transformation lies the fact that education, in its various forms, fundamentally revolves around the effective processing of information. Information processing is a forte of AI technologies. This intersection underscores a pivotal convergence, as AI can significantly augment and innovate two core functions of education: teaching and research. In both domains, the assimilation, analysis, and communication of information play pivotal roles. The application of AI in education not only holds the promise of enhancing instructional techniques and personalizing learning experiences but also revolutionizing the research landscape by enabling novel insights through advanced data analysis. This paper delves into the evolving relationship between AI and education, exploring how AI's information-processing capabilities are reshaping the educational landscape and propelling us into a new era of learning and research. (Gassmann et al., 2023)

The relevance of the EdTech sector in combination with AI is driven from both sides, explained Meta’s Head of AI Jerome Pesenti on the Podcast Weights & Biases: The AI community sees education as a potential market and the education sector sees AI as a potential new tool. (Fields & Pan, 2022, min. 35:55-40:15).

Writing, especially in the form of essays, is at the core of traditional education. A field substantially impacted by modern NLP systems. Universities need to overthink how they award degrees, as it can be reasonably expected that a substantial part of the average student workload can be automated by systems like ChatGPT. (Marche, 2022).

The abilities of these contemporary language processing systems are further illustrated by the fact that ChatGPT passed, amongst others, law school exams in blind grading of four tests, consisting of 95 multiple-choice-questions and 12 essay questions. The AI achieved a grade of C+, ergo below average but passing (Choi et al., 2023). These results raise obvious questions of fairness and justification of grades awarded to students.

In another Experiment at Wharton ChatGPT fared well too. The key takeaways include the advice of banning AI tools when testing foundational knowledge, raising awareness for its limitations, and deploying the technology to improve the productivity of the teaching process (Terwiesch, 2023).

2.2 INVESTOR BEHAVIOR

Establishing a business encompasses the pursuit of opportunities, the allocation of resources, and the creation of value. To bring their visions to fruition entrepreneurs need funding. While a business fulfills different needs of different stakeholders, this thesis focuses on its role as an investment object. The previous chapter therefore describes the target, while this chapter will discuss investments as an activity.

On one hand, the essential theories underpinning company valuations are discussed. On the other hand, the evolution of behavioral finance and investor psychology is laid out.

2.2.1 BUSINESS VALUATION AND INVESTMENT BIASES

In general, new ventures are traded on private markets (venture capital) and not on public markets (stock exchanges), thus market prices are highly dependent on the valuation of a few involved parties. In practice, business leaders mostly rely on the Discounted Cash Flow (DCF) method of valuing companies or projects. The method consists of two key components: the expected cash flows and the discount rate. It is thus future-oriented and introspective. It aims to factor in opportunity costs through the use of the discount rate, leading to the estimate of the present value of each project or enterprise (Sahlmann & Scherlis, 2009; Stevenson et al., 1999).

It is often accompanied by multiples as a secondary method, whenever data from financial markets are available the P/E multiple is the most popular one. The multiples approach is based on the known values of industry peers, thus reflecting an estimate based on external, market-based valuations. This latter method is highly dependent on the selection of companies included in the peer group. (Berndt et al., 2014; Koller et al., 2005).

Start-up valuation is more complex than valuing long-standing, publicly traded firms. This complexity is mainly driven by a lack of historical, reliable financial data from within the venture. There is no established baseline of financial performance to anchor growth projections. The valuation is thus often based on the potential and ability of the team rather than quantifiable financial data. These assumptions are then codified in business plans. A practical example features the T2D3 (triple, triple, double, double, double) growth model. These very rough growth estimates illustrate again, that valuation for start-ups is mostly about agreeing on a price range rather than precise calculations. (Cavalry Ventures, 2022). Additionally, there is a specialty in valuing AI companies that is noteworthy in the context of this thesis: Competitive advantage and thus valuation is driven by the amount and quality of data an AI start-up controls (Bessen et al., 2022).

The intersection of cognition and management research provides a rich and diverse landscape for exploring how individuals and organizations perceive and make sense of their environments. At the core of this research lies the concept of mental frameworks, serving as reference points that structure information and guide decision-making processes. These cognitive representations that individuals have developed from past experiences are used to navigate future challenges. The term dominant logic encapsulates how individuals and organizations conceptualize their industry, operationalize their past successes, and shape their decision-making processes. It essentially serves as a prevailing mindset that influences not only the strategies formulated but also the means employed to achieve goals. (Gassmann et al., 2016, pp. 50–54; Prahalad & Bettis, 1986).

The Efficient Markets theory enjoyed its heyday in the academic world during the 1970s, asserting that financial markets fully incorporate all available information, making it impossible to consistently outperform the market. However, this theory was progressively undermined by a series of discoveries in the 1980s, revealing market anomalies and evidence of unexpected volatility in returns, paving the way for the emergence of behavioral finance. Since the 1990s, behavioral finance has flourished as a subfield, focusing on understanding how psychological factors and irrational behaviors influence financial decision-making. Notable developments include feedback theories that elucidate market dynamics, models that explore the interaction between informed investors and ordinary participants, and extensive research on sophisticated investors in financial markets, also known as "smart money". This paradigm shift towards acknowledging the complexities of investor behavior and market dynamics has reshaped our understanding of financial markets and paved the way for more holistic and inclusive theories that encapsulate both rational and irrational elements in decision-making. (Shiller, 2003).

Perception biases play a central role in shaping financial decisions. A wide array of potential biases is described. The findings indicate that investment biases may be inherent expressions of deeply rooted, evolutionarily aspects of human behavior. (Heath & Tversky, 1991; Kahneman & Riepe, 1998; Kahneman & Tversky, 1979; Tversky & Kahneman, 1974, 1992)

The most relevant of these biases in the context of this thesis is the concept of conformity bias, sometimes also referred to as "Group Think". This cognitive bias potentially compels investors to react to prevailing market trends or consensus opinions, leading to herding behavior. This phenomenon has also been described as the “overextrapolation of trends” (Alti & Tetlock, 2014). The presence of conformity bias can have profound implications for financial markets, as it can drive asset prices away from their intrinsic values, contribute to asset bubbles, and amplify market volatility. (Moscovici & Faucheux, 1972).

Important to note in the context of this thesis is, that individuals who have accumulated experience in finance exhibit a reduced inclination toward these biases (Cronqvist & Siegel, 2014).

This thesis tries to understand financial decision-making, with a specific focus on trend-driven biases, analyzing their manifestation in the context of investments in the EdTech market. Understanding these biases is pivotal in improving investment strategies or advancing fundraising strategies.

2.2.2 MEASURING TRENDS AND INVESTOR BEHAVIOR

In simplified terms, this thesis investigates whether investors make good decisions or if their perception is skewed by the adoption of AI.

At the basis of my approach is the fact that text produced by firms holds meaning for the market. Through marketing material and communicative efforts, the firms make their value proposition known. While primarily these statements are addressed towards customers, investors also make use of this information. (Abrahamson & Hambrick, 1997; Guzman & Li, 2023; Hoberg & Phillips, 2016).

This stated value might significantly diver to the value investors assign a firm. The primary variable of analysis is the total funding a venture receives, it is the most direct expression and quantifiable measure of investor behavior in the dataset. It is measured in USD and indicates how attractive a business idea is to investors.

Adapting the definition of Blei et al. (2003), which suggests considering a topic “hot” in science when there is growth in the number of published contributions to a topic, for the VC context means, that a “hot” topic is delimited by the relative increase in new ventures created assigned to a topic. The same data pattern is classified as “hype” elsewhere (Buck et al., 1998, p. 65).

Combining these two ideas means first establishing topics and trends within the dataset. By analyzing the sensitivity of investors to the ventures within a "hot" topic period I establish a measure for the quality of investment decisions. This is effectuated by contrasting the average investments during and outside the "hot" phase.

The idea is best exemplified by the findings of Cooper, Dimitrov, and Rau (2001) already mentioned in the introduction: A mere affiliation with the internet through a simple name change significantly enhanced a firm's value during the late 1990s. This association, regardless of the extent of actual internet utilization in the firm's operations, triggered a substantial and enduring increase in the firm's stock price. This phenomenon underscores the power of trend-based biases regarding business valuation.

3 METHODOLOGY

Conceptually I aim to find insights in a pre-existing dataset. Traditionally management science is theory driven and subsequent research is undertaken to prove or disprove the theory. In this case, the data is used for fact-finding and theory validation in a specific sector. The following chapters shed light on the data collected and the tools used to evaluate said data. For further technical information please consult the provided iPython Notebooks (.ipynb-files) containing the original code[2].

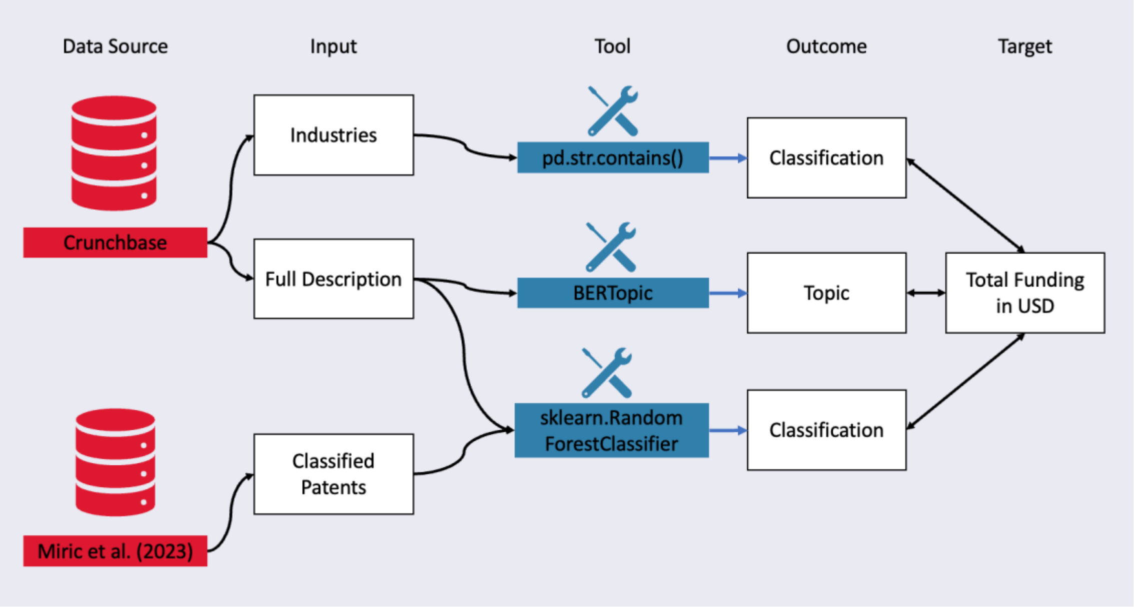

Figure 1 Research Design

In terms of methodology, this thesis employs topic modeling and classification algorithms on company descriptions to recognize their connections with AI technology. The topic modeling is done using BERTopic, thus employing a purpose-built library utilizing a modern Large Language Model (LLM). The classification follows a two-pronged approach. On one side the self-declared industry-tag is analyzed. On the other hand, Machine Learning (ML) is employed. The company

[2] See Appendix or https://github.com/MarioMuehlematter/Education.ai

descriptions are reduced to a Bag-of-Words (BoW), which in turn is then classified as containing AI-related words, word fragments or word stems. This is done following the example of Miric et al. (2023), who identified Random Forest classification (RFC) as the most promising approach for management science. With regards to data quality and completeness of the dataset at hand, total funding was focused upon. Survival and the number of employees were considered as well, though they were not treated in depth and only used to assure data integrity and quality.

3.1 DATA COLLECTION AND PRE-PROCESSING

In the following chapter, the process of collecting and subsequentially curating the necessary data is described. Additionally, some high-level insights into the obtained dataset are described. The dataset was prepared with a focus on enabling NLP later in mind.

3.1.1 DATASET INFORMATION

The dataset was extracted from CrunchBase (CB). The investor platform features an extensive collection of data that is also valuable and useful for science. The data on CB is considered reliable. It is one of the two dominant VC databases for academic research, the other major player being Pitchbook. Together they provide the foundational data for approximately 75% of VC-related publications after 2015 (Retterath & Braun, 2020, p. 7). In 2017 there were 90 articles and papers that relied on CB, most were recent, and the first one dates back to 2009 (Dalle et al., 2017, p.16).

The information contained in CB is dominated by the American start-up ecosystem, the clear majority of listed firms have less than 50 employees (FTE) and are US-based (Dalle et al., 2017, pp. 8–11).

Other known biases in CB data include a positive selection bias, as entrepreneurs with high-quality ventures have an incentive to increase their visibility and send signals to potential investors subsequently increasing the value of their shares. Despite these limitations, CB offers the best information regarding financing rounds and total funds raised (Retterath & Braun, 2020, pp. 3–4).

Dalle et al. (2017, pp.16–18) note that there is a high variety of research questions addressed by using CB data. They further identified four research clusters based on journals, the main areas of research interest of said publications were: Management, Green Innovation, Technology, and Small Business. The third group, Technology, applies to this paper.

There are also two other clusters of research, not defined by publication outlet but by topic, that are relevant to this master's thesis. The first group concerns research based on CB data that analyses investor behavior (Liang & Yuan, 2016; Zeng et al., 2016; Zhong et al., 2018). The second group consists of examples of other authors using CB and Topic Modelling (Chae & Olson, 2021; Shi et al., 2016).

The earliest founding date in the dataset underlying this master's thesis is the 1st of January 2009 and the data collection was cut off at the end of 2022. The initial dataset exported from CB contained 29,346 company profiles, with 112 columns each. In pre-processing companies whose descriptions are too short (word counts <30) to be useful for topic modeling were excluded, following the example of Chae & Olson (2021). Virtually all entries in the simple "Description" column of the dataset are too short, according to established data science practice. Therefore the "Full Description" column was used for further processing. 14,046 entries and 1,225,060 words remained in the resulting dataset used for analysis. The company descriptions therein contain 87.2 words on average.

Each company profile of the final dataset had 112 columns, ranging from "Organization Name URL" to "Aberdeen - IT Spend Currency (in USD)", the completeness of these entries varies widely, a full list of column names is available in the appendix. Naturally not all 112 fields were considered in the analysis, I actively used the following information: Full Description, Founded Date, Headquarters Location, Number of Employees, Total Funding Amount Currency (in USD), and Operating Status.

Lastly, a high-level insight: the mean total funding amount of the ventures contained in the dataset was 16,354,648 USD.

3.1.2 DATA INTEGRITY AND QUALITY

Previously I established that CB is considered a relatively reliable data source. It was also noted that certain biases are usually present. These established biases were controlled for in this thesis.

Firstly, when analyzing the number of ventures per country of origin, the focus on the US start-up scene is confirmed (see Fig. 2). It is noteworthy, that the top 10 countries of origin also contain relatively small countries like Switzerland, The Netherlands and Israel, all nations are also considered very innovative by the World Intellectual Property Organization (WIPO, 2022). However, other very innovative countries (e.g. Sweden) are missing from the most prevalent countries of origin in the dataset. Striking is the contrast between the two most populous countries on earth, whilst almost naturally both China and India make the top 10, India is home to 1951 ventures while China only accommodates 314 firms of the dataset. The data offers no direct explanation of this discrepancy, it is plausible that Chinese entrepreneurs might simply favor different data providers.

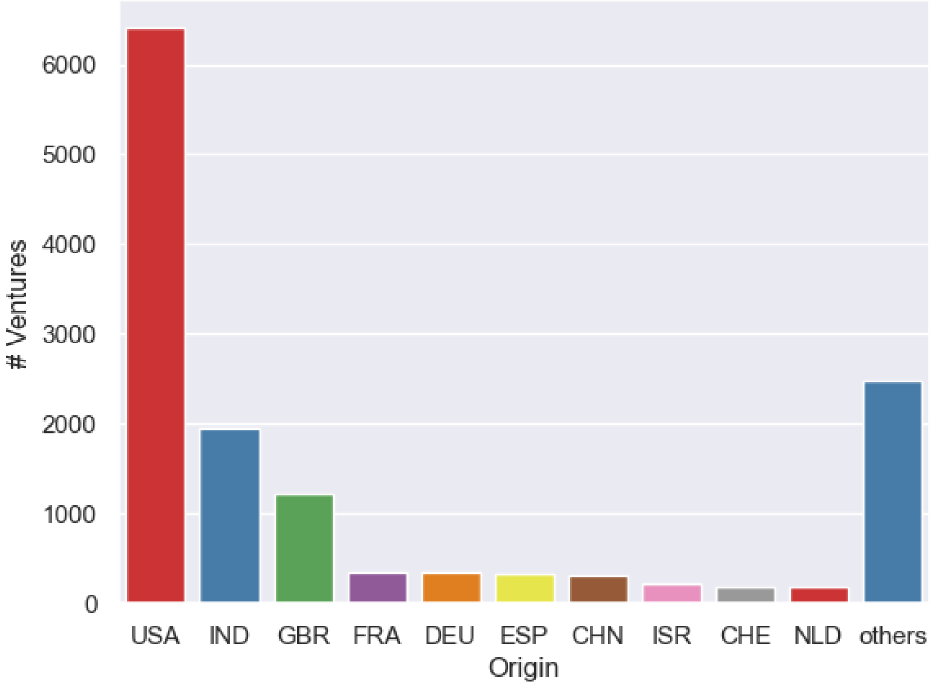

Figure 2 Number of Ventures per Country of Origin (Top 10)

(Note: 14004 out of 14046 entries contained information on location)

Furthermore, the British (GBR) are either outcompeting their European counterparts of France, Germany, and Spain (FRA, GER, ESP) by 4x or there is a bias due to language or regional preference of VC data provider.

Another insight from the literature could be confirmed for the dataset: 87.6% of ventures had 50 or fewer employees (Note: 13248 out of 14046 entries contained information on headcount).

In conclusion, these outputs indicate that the findings of Dalle et al. (2017) are upheld for this EdTech-focused dataset. Company data provided by the CB database remain mostly US-based and small (<50 FTE).

The survival rate, as a secondary indicator, reflects the particularities of the start-up environment. While failures of established, publicly traded corporations are relatively rare, start-ups fail often. (Bureau of Labor Statistics, 2023; Gyeonggi Northern Chamber of Commerce and Industry, 2021). It makes therefore sense to measure the survival rate. Higher survivability does not directly indicate that a certain group of start-ups is more profitable but could also be driven by a higher willingness of investors to finance the losses[3] of a group. To ensure optimal transferability and enable an insightful comparison this thesis appropriates the OECD definition of survival rates: *"The number of n-year survival enterprises for a particular year t refers to the number of enterprises which had at least one employee for the first time in year t-n and remained active in year t. This definition of survival excludes cases in which enterprises merge or are taken over by an existing enterprise in year t-n. The employer enterprise survival rate in sector (class size) x measures the number of enterprises of a specific birth cohort in sector (size class) x that have survived over different years. the n-year employer enterprise survival rate for a reference year t is calculated as the number of n-year survival enterprises as a percentage of all enterprises that reported at least one employee for the first time in year t-n. the share of n-year-old employer enterprises for a particular year t refers to the number of n-year survival enterprises as a percentage of the total employer enterprise population in year t. "* (OECD, 2017, p. 88). The survival rate offers a valuable checkpoint for data quality and reliability as it enables a comparison between the dataset and the wider economy. Retterath and Braun (2020, p.3) note that "VC databases should be more likely to collect self-reported information on higher-quality firms, resulting in a positive selection bias". To control for said bias, additional data from the OECD and the USA, as it is the most common country of origin in the dataset, is introduced. The first is provided by Gyeonggi Northern Chamber of Commerce and Industry (2021), the latter sourced from the Bureau of Labor Statistics (2023). The additional data reveals, that the positive selection bias suggested in Reterath & Braun (2020) is present (see Fig 3). The discrepancy in the survival rate of firms is stark from the beginning to the end. This finding indicates that the dataset is not fully representative of the wider American or Western economy.

[3] Start-ups often use the terms cash-burn and runway, cash-burn being the operational loss per period and runway being the time until new financing is necessary or else the company defaults.

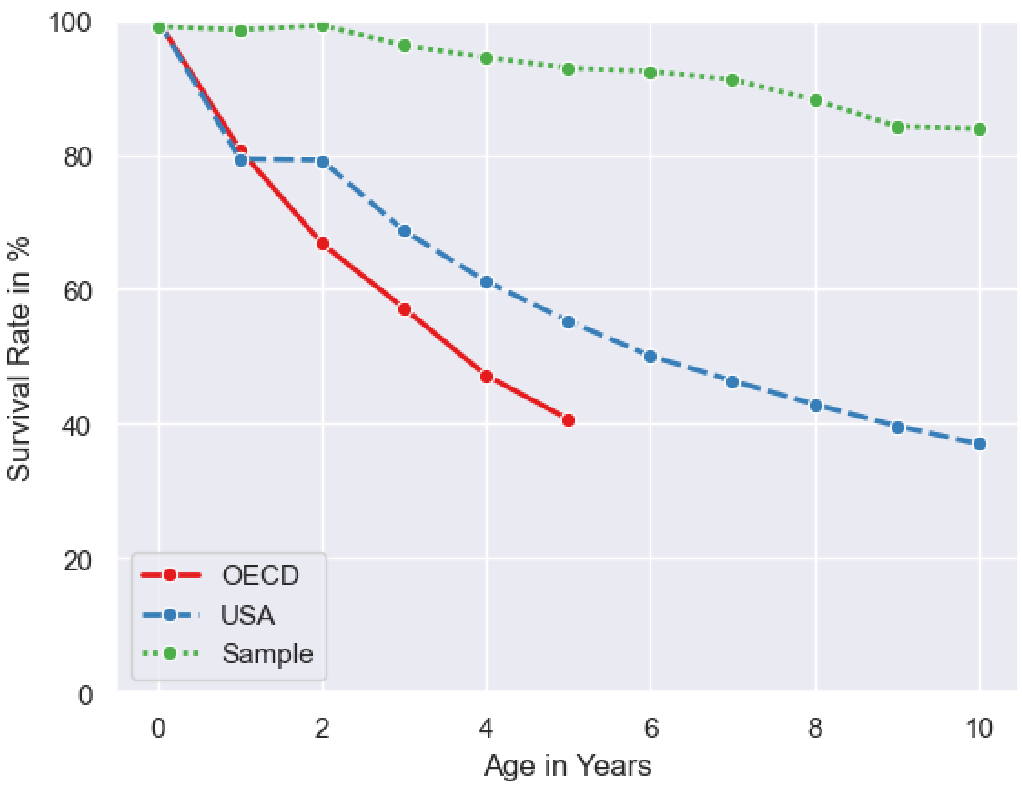

Figure 3 Survival Rate Comparison

Additionally, the US survival rates show the "COVID-Freeze" in bankruptcy, whilst the OECD values used as baseline are older and do not feature the pandemic bump, caused by government stimuli. The pandemic might also partially explain the ultra-high survival rate in the EdTech dataset.

3.2 DATA PROCESSING

The following chapter recounts the steps involved in processing the data with two different methods of Natural Language Processing (NLP). By leveraging the strengths of each method, namely Topic Modeling (TM) and classification enabled by Machine Learning (ML), the understanding of the data was deepened. The aim was to uncover latent patterns, relationships, and classifications that might remain elusive when relying on a single technique. This dual-method approach not only contributes to the comprehensiveness and robustness of my analysis but also highlights the versatility of NLP as an indispensable tool in the realm of data-driven research. Working with natural language data requires a common understanding of the following terms:

- *A word is the basic unit of discrete data, defined to be an item from a vocabulary.*
- *A document is a sequence of words.*
- *A corpus is a collection of documents.*

The definitions above are adopted from Blei et al. (2003, p. 995), with only minor simplifications.

3.2.1 STEP 1: TOPIC MODELING (BERTOPIC)

The first analysis in this thesis uses TM in conjunction with the founding date of each firm to identify "hot" or "cold" topics. Griffiths and Steyvers (2004) define, for an academic setting, topics gaining prominence based on the increasing number of publications during a period as "hot", the inverse is considered "cold".

This idea holds substantial potential for adaptation within the context of start-up ecosystems. This is achieved by substituting the count of scholarly articles with the number of newly founded firms self-identifying with specific topics. In this context, "hot" topics signify areas of entrepreneurial interest and activity experiencing growth and innovation, while "cold" topics represent the converse, characterized by waning entrepreneurial engagement. By leveraging this concept, valuable insights into the evolution of the start-up landscape are gained. Helping to identify areas of opportunity, anticipate market shifts, and guide strategic decision-making.

In 2003 a paper established the origins of TM in the modern computational sense, making use of the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). However, technology has evolved since then. Based on transformer technology Large Language Models (LLM) have emerged (Vaswani et al., 2017). These models enable more sophisticated algorithmic text analysis.

Currently, there is a wide range of LLMs, based on different model architectures and training data, available for research purposes. The best-known architectures include GPT, deployed by OpenAI in their own ChatGPT as well as Microsoft's Bing AI, Llama researched and published by Meta, and PaLM powering Google's Bard AI. (Thompson, 2021).

Whilst being the sub-part subject to the most dynamic change at the moment, the TM methodology used in this thesis consists of more than an LLM. To secure a stable and reliable outcome the BERTopic Python library was used (Grootendorst, 2022). The architecture of said analysis tool is modular. The tech stack, meaning the modules utilized in this thesis were as follows:

This thesis makes use of the language representation model BERT (standing for Bidirectional Encoder Representations from Transformers). BERT is able to successfully tackle a broad range of NLP tasks, whilst remaining conceptually relatively simple and comprehensible (Devlin et al., 2019). Derivatives of the original BERT, remain at the cutting edge of technological development, measured by the GLUE benchmark (gluebenchmark.com, 2023; Wang et al., 2018). The exact variant used in the tech stack for TM is called SBERT (Reimers & Gurevych, 2019). This model is used to transform the text data into a mathematical vector representation.

The corpus-level multi-dimensional embeddings created using SBERT, are subsequentially reduced using UMAP (standing for Uniform Manifold Approximation and Projection). This non-linear technique is distinguished by its ease of use and speed (McInnes et al., 2018).

Once the embeddings have been reduced, the next critical step in data analysis involves clustering. To achieve this, the power of HDBSCAN, a density-based clustering technique, was harnessed. Unlike traditional clustering methods, it is adept at identifying clusters of various shapes and excels in pinpointing outliers within the data. Consequently, this approach promises to enhance the quality of our resultant topic representation by minimizing noise. By allowing the data to selforganize into clusters based on density, HDBSCAN facilitates a more accurate and nuanced classification of documents. (McInnes et al., 2017).

After the documents are clustered, another proven technology, which also underpins modern Internet search engines, is applied to the corpus. Taking regress to an idea originally proposed by Spärck Jones (1972), TF-IDF (standing for term frequency-inverse document frequency) measures the importance of a word to a document in a corpus, adjusted for the fact that some words appear more frequently in general. TF-IDF is a well-established method for assessing word importance across a set of documents. It provides valuable insights into the relative significance of words in the context of the entire corpus. (Blei et al., 2003, pp. 993–994).

The BERTopic library uses this established approach innovatively by treating all documents within a specific cluster as a single corpus. By doing so and then applying TF-IDF, importance scores for words within that cluster are obtained. The rationale behind this method is that the more important a word is within a cluster, the more representative it is of the underlying topic encapsulated by that cluster. In essence, this approach enables the extraction of the most representative words per cluster, yielding concise and descriptive topics. This novel TF-IDF application distills meaningful information and topic descriptions from complex document collections. (Grootendorst, 2022).

To summarize the process at a high level: So-called tokens (word fragments) are extracted from the analyzed text. Words are turned into number representations in that way and the resulting vectors are then compared to each other to build clusters. Reversing these clusters back to words, topics become evident. Once the topics are assigned to each dataset entry, adding the temporal component to the analysis is simple, as the date variable present in the original dataset can be used.

3.2.2 STEP 2: CLASSIFICATION USING SUPERVISED LEARNING

The second Natural Language Processing (NLP) technique employed in analyzing the dataset for this thesis was a form of supervised machine learning. This classification approach promised to offer a deeper and more nuanced perspective on the complex phenomena under investigation. This technology focuses on word-level analysis. It tries to extract informative fragments from the tokenized text, thus is a different approach to NLP entirely in comparison with the methods described in chpt. 3.2.1.

Following an existing example (Miric et al., 2023), the process of applying a Random Forest classifier (RFC) consists of multiple steps, as it is a form of supervised machine learning the model needs training before applying it. The training consists of tokenizing the text, then vectorizing these tokens, and assigning the known binary classification to the corresponding vector spaces. When applying the previously trained classifier to the CB data the tokenization and vectorization steps are repeated, and the binary classification is inferred based on the training dataset.

The tokenizing step is done using existing utilities contained in the Python-libraries nltk (Bird et al., 2009) and Genism (Řehůřek & Sojka, 2010). The process of extracting tokens from words results in a Bag of Words (BoW) containing word fragments. In the same process step, so-called stopwords (a collection of words that usually contain no meaning, e.g. the words "is" and "the") are removed.

The BoW is then vectorized, ergo transformed into a numeric representation, by applying the "TfidfVectorizer" contained in the Scikit-learn library (Pedregosa et al., 2011). The Python implementation of the untrained RFC model, originally conceived by Breiman (2001), can also be found in the same library.

The Random Forest algorithm is a versatile meta-estimator. It operates by constructing an ensemble of decision tree classifiers on diverse sub-samples of the dataset. The fundamental principle guiding this approach is that by amalgamating the outputs of multiple decision trees, the predictive accuracy of the model can be enhanced while simultaneously mitigating the risk of overfitting, a common challenge in machine learning. It ensures robust classification outcomes. (Pedregosa et al., 2011).

The final training step is passing the pre-processed training data, consisting of 4000 patent descriptions classified binary (containing AI or not), which was published by Miric et al. (2023). For this thesis, each tree in the RFC is trained using the entirety of the patent dataset, ensuring that no information is discarded or overlooked.

Lastly, the now-trained RFC is applied to the EdTech dataset in question, resulting in a binary classification (AI or non-AI).

3.3 DATA ANALYSIS AND VISUALIZATION

Some of the columns deemed relevant in chpt. 3.1, were manipulated to enable a clear analysis. The information fields "Age" and "Age in days" were created based on "Founded Date", and the field "Country" was derived from "Headquarters Location". The "Country" field was further converted to 3 letter acronyms, following the international standard retrieved from the ISO (n.d.).

3.3.1 ADDITIONAL CONCEPTS UTILIZED IN ANALYSIS

To uncover the trends ("hot" or "cold" topics) central to the research question a statistical tool was employed. The kernel density estimate (KDE) is a technique in statistical analysis and data visualization, that offers a means to transform discrete data points into continuous estimates. This method produces a representation of data points as a smooth, continuous probability distribution. By doing so, a more nuanced understanding of the underlying data patterns is possible. KDE accomplishes this by placing a kernel at each data point and then summing these kernels to create a continuous probability density function. This transformation provides a richer visualization of data, facilitating the identification of peaks and overall data distribution characteristics. This thesis is a practical application of kernel density estimation for dealing with discrete data points. (Parzen, 1962; M. Rosenblatt, 1956; Scott, 1979).

The selected and transformed variables in this thesis were cross-evaluated by utilizing a correlation matrix. A correlation matrix provides a comprehensive view of how variables are interrelated, quantifying the degree and direction of these associations. Potential patterns, dependencies, and areas for further investigation are uncovered by displaying correlations as a matrix. This approach enables a deeper understanding of the dataset's structure. (Rencher & Schaalje, 2008, pp. 77–78).

3.3.2 VISUALIZATION TOOLS

The analysis and visualization in this thesis were executed using the Python libraries Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007). These tools are indispensable for data analysis and presentation, offering a rich array of functions and capabilities. Seaborn (Waskom, 2021), built on top of Matplotlib (Hunter, 2007), simplifies the creation of informative and aesthetically pleasing visualizations. Its diverse plot styles make it ideal for exploratory data analysis, while Matplotlib provides a versatile platform for fine-tuning and customizing visualizations. Together, they offer a robust framework for researchers to dissect and communicate complex data patterns effectively.

4 RESULTS

This chapter lists the results obtained by the methodology described in the previous chapter and relationships to pre-existing facts are established.

4.1 AI ADOPTION IN THE EDTECH SECTOR

4.1.1 DEPLOYMENT OF AI DOES NOT FOLLOW A TEMPORAL TREND

Intended to reveal patterns in the adoption of AI technology within the EdTech market, the classification analysis of the data also sheds light on data quality aspects (see Fig. 8). It reveals intriguing patterns, particularly the striking coincidence of entry dates with the new year raises concerns regarding data quality. However, for the most recent years, the concentration around the new year is less pronounced. This improvement in data quality, suggests a maturation of data collection processes or an increased commitment to accuracy within the industry. This positive trajectory bolsters confidence in the reliability of the dataset, despite the initial temporal clustering.

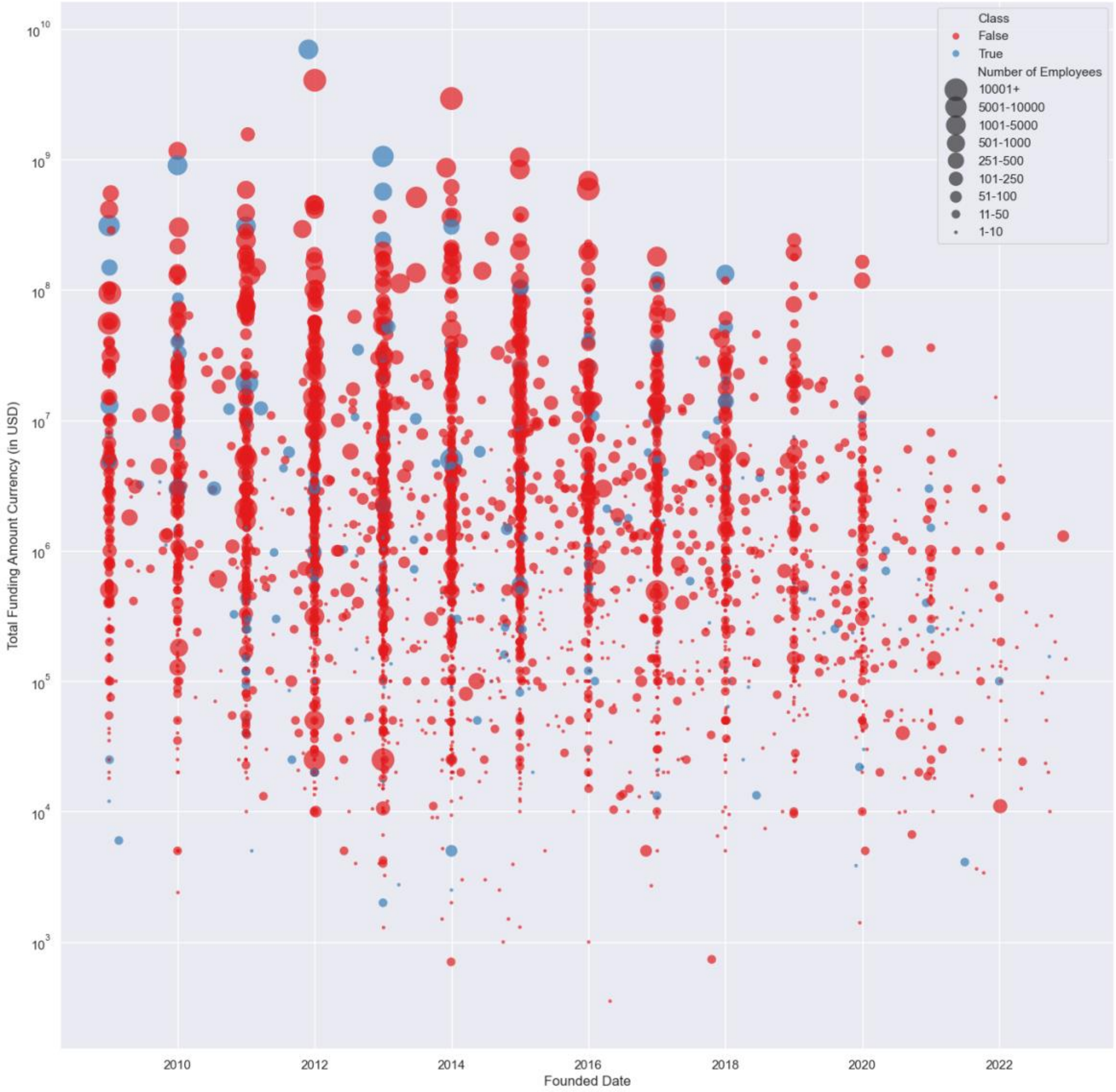

Figure 4 Scatterplot Highlighting the ML-based Classification.

Furthermore, the analysis unveils an expected and unsurprising correlation between an organization's age and total funding. Large, well-funded entities consistently gravitate toward the upper-left quadrant of the classification, reinforcing that financial stability and organizational longevity are intrinsically linked in the competitive landscape.

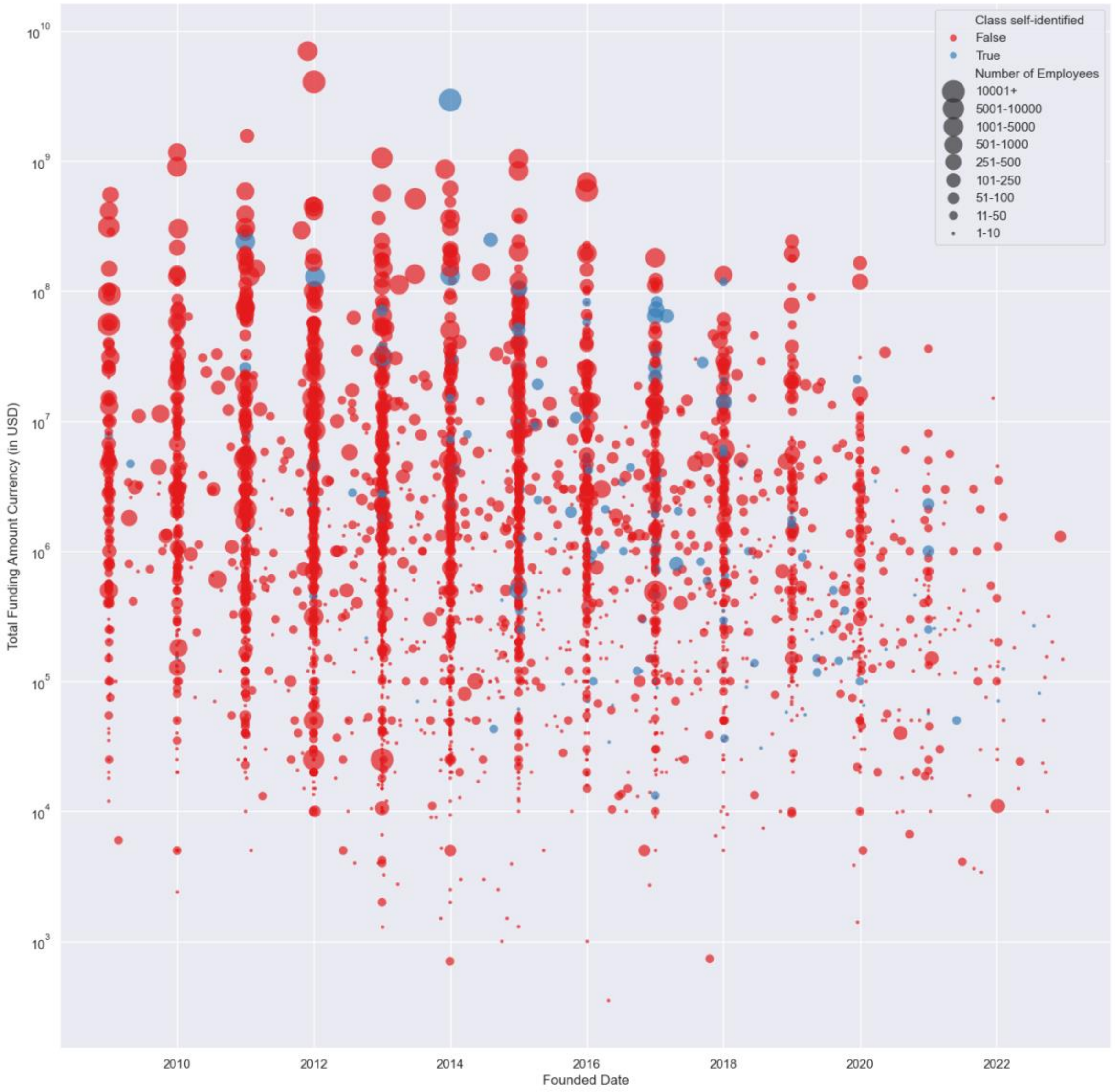

Figure 5 Scatterplot Highlighting the Self-identified Classification.

Returning to the original focus, when exploring the spatial distribution of AI companies, the pattern appears random, without any visible clustering. This is true for both ML-based classification (see Fig. 4) and for self-identified classification (see Fig. 5). The RFC identified 1070 AI companies in the dataset, or approximately 7.6% of all firms. The dataset featured 496 self-identified AI companies, rounding out to 3.5% of all companies. In both cases, all the 14,046 entries in the dataset were assigned a classification.

The randomness of the spatial distribution of AI companies is proof that AI is currently not a "hot" topic in EdTech. Thus, RQ1 is rejected.

4.1.2 AI DOES NOT EMERGE AS A STAND-ALONE TOPIC

Topic Modeling using the BERTopic code library yielded 23 topics. 8352 entries out of the sample of 14046 companies could not be assigned to a topic and were thus excluded from further analysis. In most cases, their description was not specific enough (visible in the code in the appendix as the topic "-1_ learning_students_platform_education"). These descriptions used mostly vague umbrella terms, like learning, students, platform, and education, that are commonly used throughout the entire corpus of company descriptions, thus they hold no relevant information for clustering. These words identify the company as a participant in the EdTech sector, they would hold meaning when clustering a more general sample of businesses but are useless when aiming to drill deeper within the sector. The size of a cluster or topic was set to a minimum of 100, resulting in a manageable number of topics for further evaluation. This in return also constitutes a reason for exclusion as smaller clusters were disregarded. The size of the topics ranged from 698 instances of "children_kids_parents_games" to 103 instances of "security_cyber_cybersecurity_services".

Counter to the research question formulated at the beginning of this thesis, the topic of "Artificial Intelligence" did not occur, nor did any closely related verbiage, such as "Machine Learning" or "Large Language Model", constitute a topic of its own. The absence of AI as a topic when clustering EdTech firms contained in a CB dataset suggests that entrepreneurs do not think of AI as a standalone business opportunity in the education space, certainly not when describing their companies. With regards to the research question, it is thus neither a "hot" nor a "cold" topic, as it is not an identifiable topic at all.

The information contained in the dataset also allowed for a temporal analysis of the topics. Using Kernel Density Estimation (KDE), the relative prevalence of a topic over time was visualized (see Fig. 5). This resulted in multiple leads that might constitute concrete insights when further drilled down.

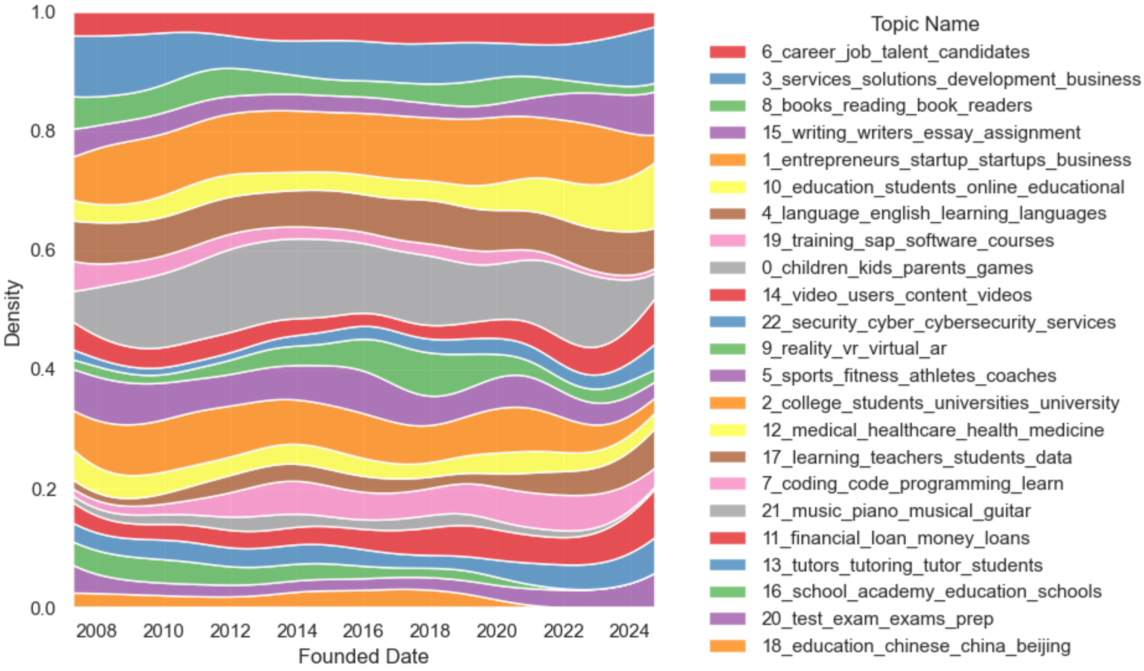

Figure 6 Kernel Density Estimate Indicating the Relevance of Topics Over Time.

Upon closer inspection, the trend estimates revealed one area of interest related to the adoption of AI. The promising anomaly is the growing number of ventures within the topic “15_writing_writers_essay_assignment” created after 2020. The increasing number of start-ups labeled “writing_writers_essay_assignment” in that period might be a proxy for LLMs as the technology rose to prominence for its text processing capabilities.

Given the insights based on KDE, subsampling topic 15 whilst also limiting the temporal scope of data to the years 2021 and 2022 seems promising. This temporal limitation coincides with the start-up definition (firm <2 years), thus general insights about start-ups can be found in the same analysis step.

The temporal limitations left 511 entries in the subsample, 9 thereof were assigned to topic number 15. Subsequent analysis revealed no heightened self-identification with the industry tag "Artificial Intelligence" in comparison with the overall dataset nor was the RFC AI tag more prevalent in this subsample. Therefore, Topic 15 cannot be considered a proxy for AI or LLM technology. Underlining again, that AI, even when considering proxy terminology is not a stand-alone topic in the EdTech market. Logically it can thus neither be "hot" nor "cold".

Furthermore, the companies in topic 15 only amassed a mean funding of 24,182 USD. They underperformed both the whole dataset, coming in at 16,354,648 USD, which is to be expected as the ventures contained in the dataset are overall more mature, and their start-up peer group averaging 1,191,062 USD. The association with the topic "15_writing_writers_essay_assignment" proved not to be of advantage in fundraising.

The best-funded EdTech start-ups in the start-up subsample are focusing on university students, the only other start-up topic that attracted more than 2 million USD in funding on average is the topic "6_career_job_talent_candidates" encapsulating Human Resources related words (See Fig. 6).

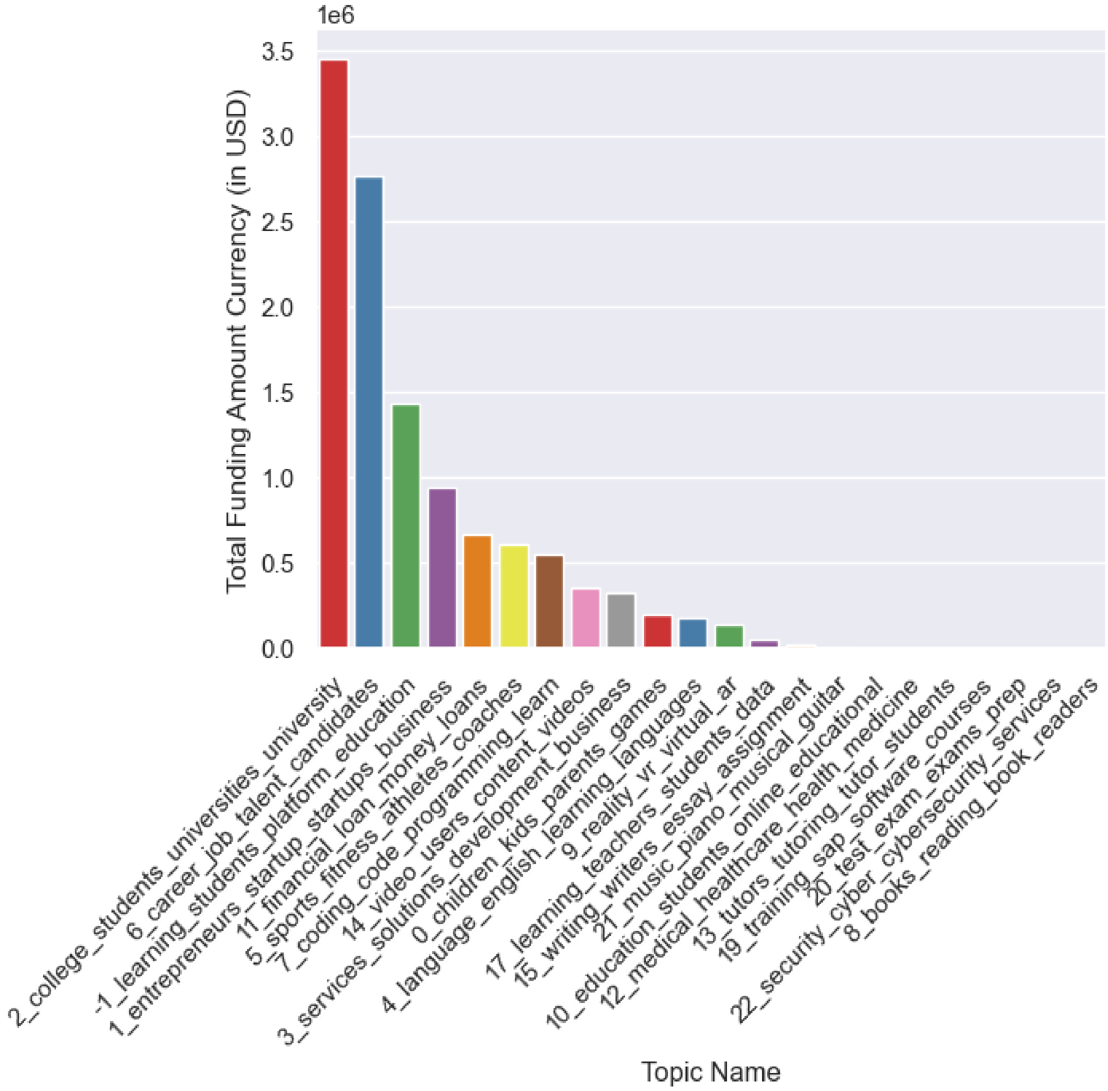

Figure 7 Subsampling for Ventures (< 2 years)

In the final analysis step, the correlation between the NLP-based TM, the self-identified classification, and the ML-driven classification and the target variable of the total funding amount in USD is evaluated.

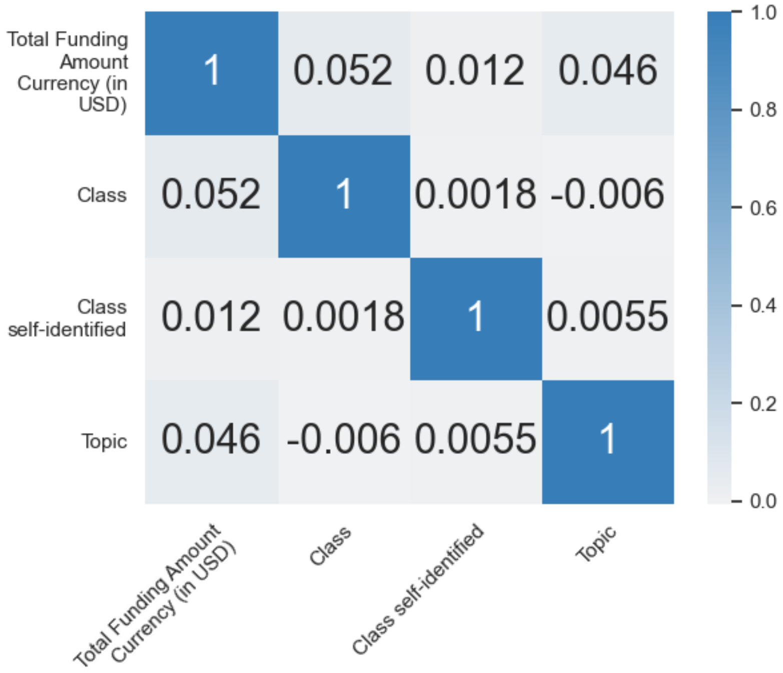

Figure 8 Correlation Matrix

The correlation matrix (Fig. 8) reveals a very low correlation for all analyzed output variables. This leads to the conclusion that in the EdTech sector based on the analyzed dataset, the association with AI does not correlate with successful fundraising. Thus, it can be concluded that association with AI does not in fundraising. Supporting the answer to research question one, that AI is not the "hot" topic in the EdTech sector, neither in terms of the number of firms established nor in funding allocation.

4.2 EMERGING TECHNOLOGIES IN EDTECH

The application of TM to the EdTech dataset yielded other valuable insights, not connected to AI. Based on KDE (see Fig. 6) topic 9, encapsulating verbiage connected to Virtual Reality (VR), offers a clear link to emerging technology. The firms associated with this topic might provide evidence of past "hot" topics in the EdTech sector.

4.2.1 TOPIC 9: VR – HOT TOPIC OF THE PAST

The interest in the wider digital economy outside the dataset surrounding VR and the metaverse peaked in October 2021, when The Facebook Company renamed itself to Meta (Meta, 2021). However based on the KDE analysis and following Blei, Ng, Michael, and Lafferty's (2003) definition VR should be considered a "hot" topic during the years 2016-2019.

All VR ventures in the dataset received a mean total amount of funding of 7,531,751 USD, this is more than VR firms amassed during their "hot" phase where their relative number soared. The mean of 6,572,573 USD during the VR heydays is also less than the overall mean funding amount of 8,564,851 USD across the entire dataset in the same period (See Fig. 9). It appears that in the EdTech Market, an association with "hot" topics had a negative impact on fundraising in the past.

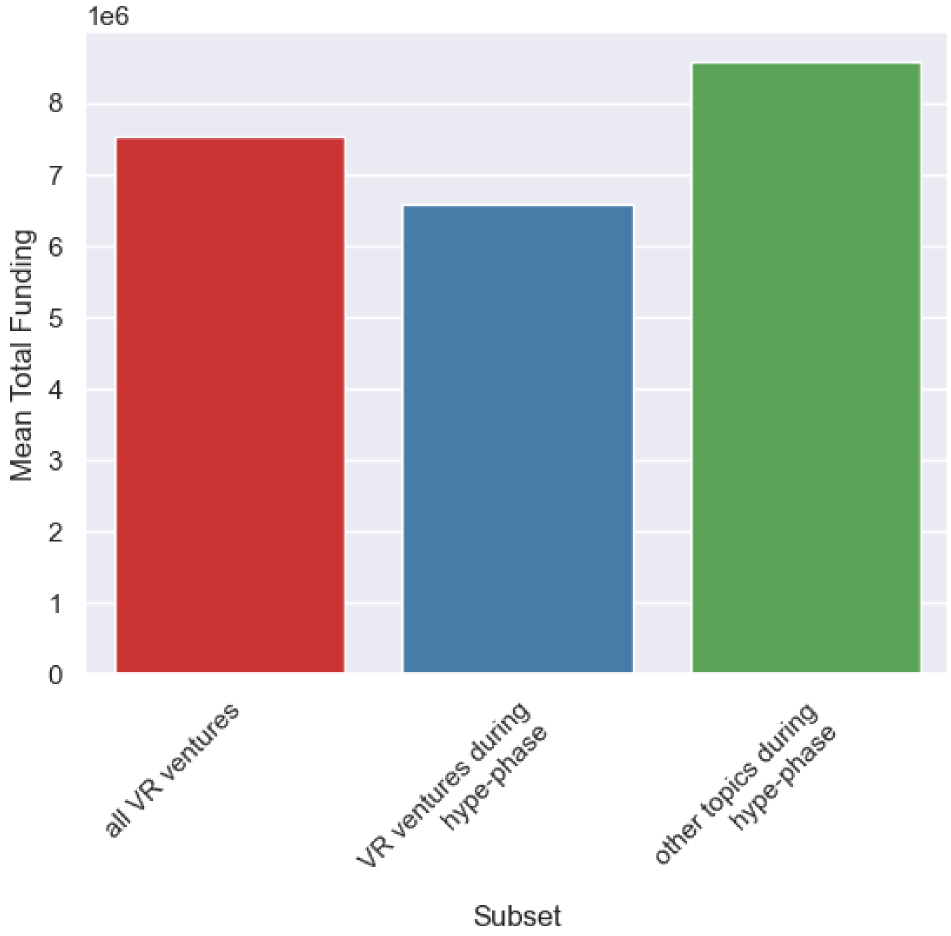

Figure 9 Mean Total Funding (in m USD) for VR Subsamples

This finding is in contrast with Cooper et. al (2001) who found a positive correlation between "hot" topics in technology and company valuation in US stock markets during the dot-com bubble. The answer to RQ2 is thus: The association with "hot" topics in EdTech does influence investor behavior. Its effect is negative on the amount of funding allocated.

4.2.2 TOPIC 18 AND TOPIC 16: CHINESE EARLY EDUCATION

The key variable used in this thesis to evaluate investor behavior is the total amount of funding received. The key variable is thus mapped and averaged to each topic (see Fig. 10). Even though the illustration uses a log scale for funding amounts topic 18 "education_chinese_china_beijing" is by far the best-funded topic.

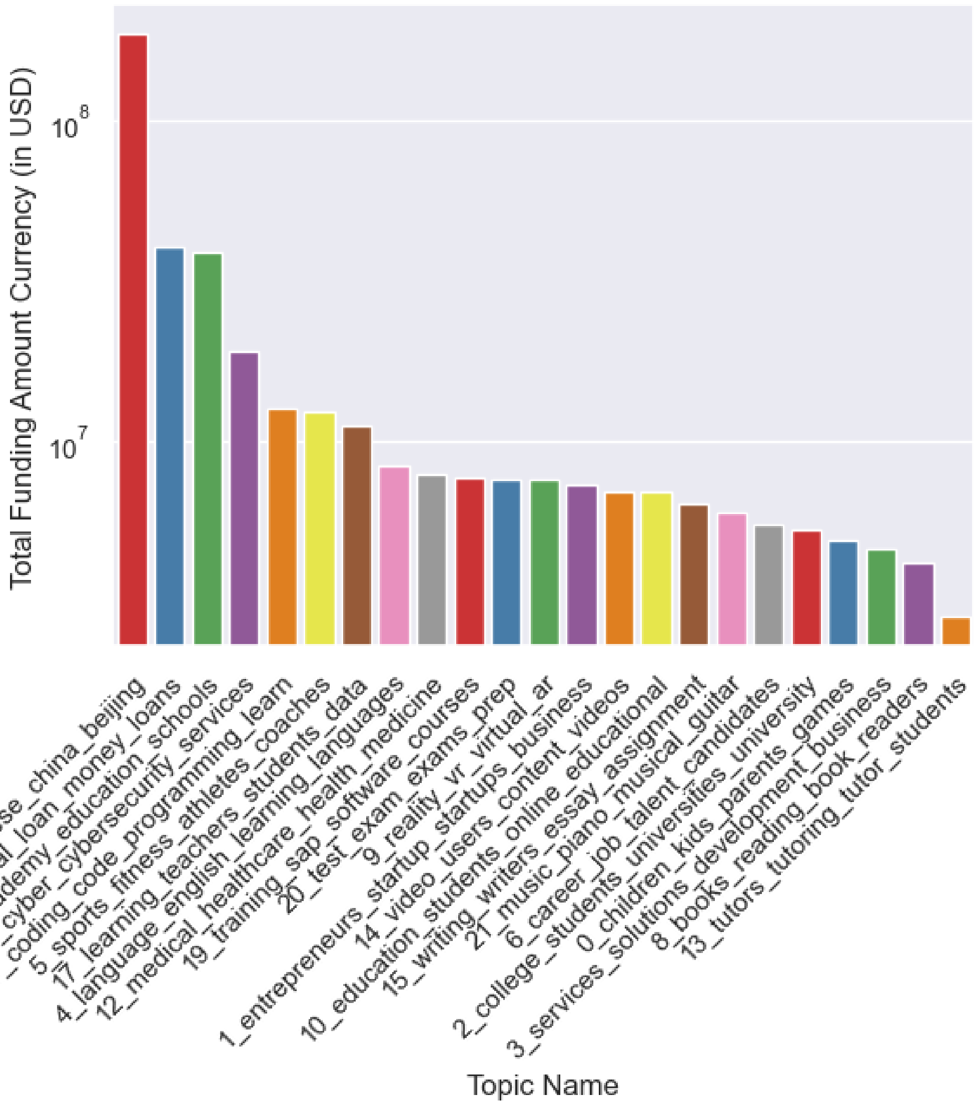

Figure 10 Mean Total Funding per Topic.

Topic 18, the best-funded topic in the overall dataset was given up by founders in 2021 and is virtually inexistent thereafter. Based on the KDE (See Fig. 6) topic 16 seems to move in lockstep.

This striking observation is driven by regulation. The city of Beijing was a leader in banning tutoring for school-aged children, as the pressure put on by parents was seen as too taxing. Later in July 2021 the rest of China followed the example of its capital region. With the ban on tutoring digital offerings in the sector were made obsolete. (Feng, 2022a, 2022b). This regulatory change in China explains the dramatic drop-off in relevance for topic 18 as a clear majority of 103 out of 137 firms are headquartered there and a further 6 firms have their HQ in the Chinese-speaking territories of Taiwan, Hong Kong, and Singapore.

Whilst the explanation for topic 18 is concise and clear, the opposite is true for topic 16. When analyzing the country of origin no striking pattern emerges, 101 of the 147 companies of topic 16 have their HQ in the USA, which can be considered inconspicuous given the remarks regarding data integrity and quality (see chpt. 3.1.2). This leaves two possible explanations: Either the firms served the Chinese tutoring market from outside the country, or the overlap is coincidental.

4.2.3 FUND RETURNERS

To gain additional insights into the investor behavior within the EdTech market the top 10 best-funded firms were extracted from the dataset (see Fig. 11). VC-Investors are always on the hunt for fund returners, the select few firms performing so well, that they finance the losses of all other investments. In the best case, they correctly identify unicorns, defined by a valuation surpassing 1 billion USD (CB Insights, 2023). BYJU'S, the best-funded EdTech venture, is a learning app from India. Notably, half of the top 10 were not assigned a specific topic and four were in the "Chinese Early Education" cluster described in the previous paragraph. Indicating that the dominating generic learning apps and tutoring apps were the big investment opportunities. According to the insights contained in this dataset, geographically these opportunities were concentrated in emerging economies, especially China and India.

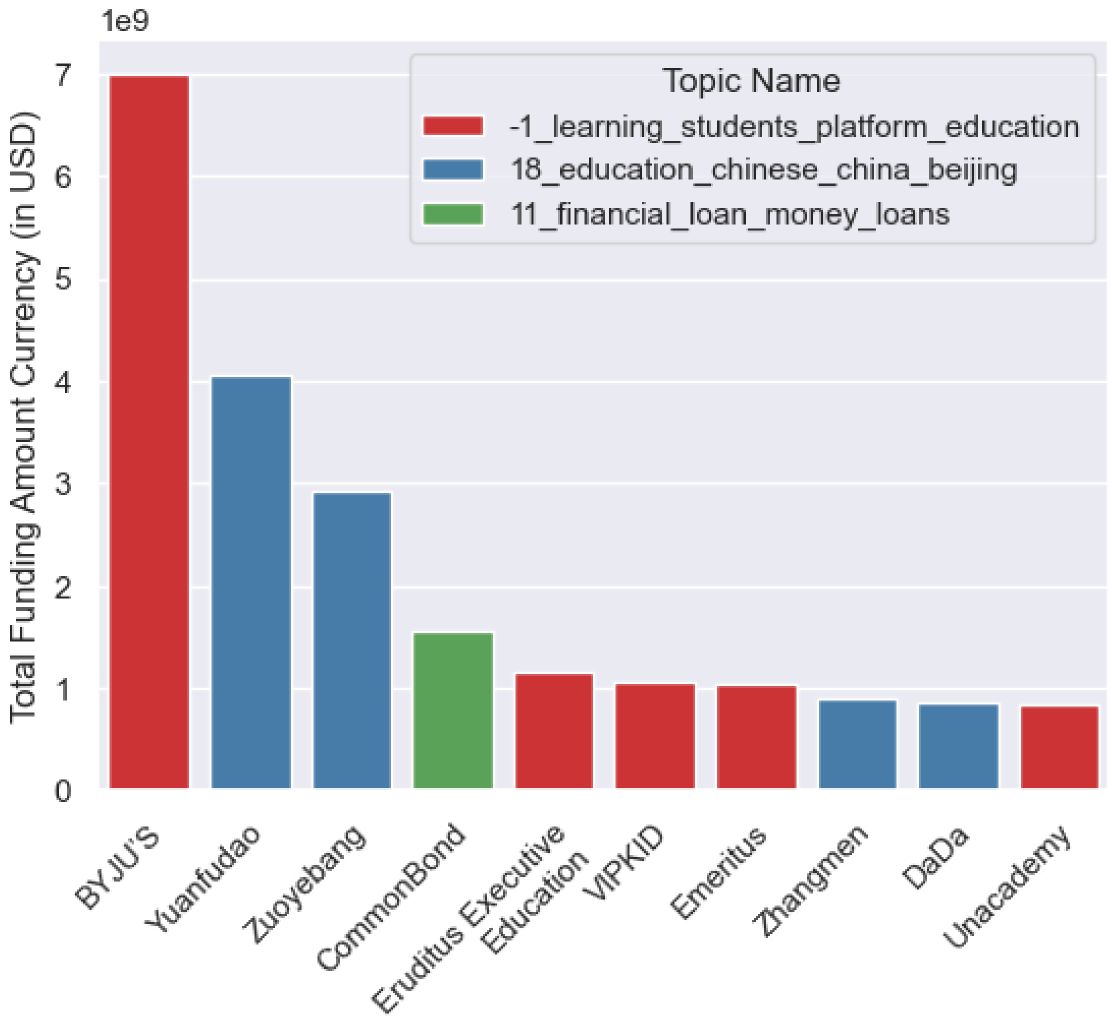

Figure 11 Fund Returners (Top10 Companies by Total Funding)

5 DISCUSSION

The subsequent chapter contextualizes the results of the NLP-based analysis of the dataset. Comparing the newly gained insights with pre-existing literature.

5.1 AI VENTURES IN EDTECH

The absence of AI as a topic when clustering EdTech firms contained in a CB dataset ranging from 2009-2022 suggests that entrepreneurs do not think of AI as a stand-alone business opportunity in the education space, certainly not when describing their companies. With regards to the research question, it is thus neither a "hot" nor a "cold" topic, as it is not an identifiable topic at all. AI is probably best seen as a tool. Referring to the four dimensions of the recombination school (Gassmann et al., 2014, 2016), it is a potential answer to the question of "WHY" or "HOW", explaining the mechanisms within a business to achieve its goals. However, AI is not a common answer to the question of "WHAT" is the core of the offering or to the question of "WHO" is the target customer. But it is exactly the question of "WHAT" and "WHO" that are most pronounced in the company descriptions, given the overall verbiage of the labels.

A second plausible explanation for the absence of an identifiable AI cluster is that innovation in EdTech might be happening outside the traditional start-up space. This insight into the deployment of AI tools by established players is evident in other industries. For example, in the mining sector, the established player Volvo is using AI to automate mines. Further prominent examples include the deployment of AI in existing search tools by Google and Bing. The most clear-cut example of this line of thought is the integration of AI in Microsoft's Office products, a business segment that was introduced in the late 1980s. Indeed, there is anecdotal evidence that established institutions are driving the integration of AI into the education sector. Harvard, for example, announced that its course CS50 will make use of AI-tutoring (Schisgall & Hamid, 2023). To further sharpen and confirm this insight a study focusing on the integration of AI technology at higher education institutions is called for.

Lastly, there is also the possibility of a time lag of AI technology penetrating the education industry. The dataset might, simply put, only cover a period too early to contain a substantial number of AI-EdTech firms that would form an identifiable cluster. A repetition of the presented work at a later point in time is needed to confirm this hypothesis.

5.2 INVESTORS ARE NOT BLINDED BY HOT TOPICS

The idea of “hot” and “cold” topics applies to the EdTech market and corresponding phenomena are visible in the data analysis undertaken. The answer to the second research question thus is: The effect of the most striking “hot” topic “9_reality_vr_virtual_ar” on funding is negative. This evidence runs counter to the findings for public markets in the USA during the dot-com bubble, which inspired this thesis (Cooper et al., 2001).

A plausible explanation, as to why the effect is inversed is the idea of “smart money”. In the realm of finance and investment, the concept of "smart money" has gained prominence, reflecting a discerning approach to capital allocation. The market structure for private markets represented by the CB dataset is unlike the frenzied and speculative nature of the dot-com bubble in public markets. There is a distinction in investment behavior and strategies as the private market is marked by a more cautious and informed approach, with a focus on sustainable growth and value creation. Furthermore, the VC landscape, which plays a pivotal role in funding firms listed on CB, is predominantly staffed with professional investors. Institutional venture capitalists, family offices, and strategic investors favor due diligence and risk assessment over impulsive speculation. As noted in the theoretical background: work experience within the financial industry reduces susceptibility to investment biases (Cronqvist & Siegel, 2014). This finding seems to be confirmed. Additionally, this group of investors tries to find exceptional companies, an aim that runs somewhat counter to following trends.

However, this thesis should not be interpreted as proof of the “smart money” concept. It is only evidence of the absence of trend-related bias in the specific and limited private market for EdTech.

6 CONCLUSION

In conclusion, this comprehensive analysis of a dataset containing 14,046 EdTech businesses has provided significant insights into the role of AI in this sector. Despite the prevalent discourse around AI as a transformative force in various industries, it appears that in the EdTech market, AI is not a primary focus but rather a tool employed to realize the visions of these companies. This observation is grounded in the absence of AI as a labeled topic among the 23 identified through topic modeling. The binary classification added in the second analytical step reaffirms that AI is not the current "hot" topic in EdTech. The distribution of AI companies concerning size, total funding, and age appears random, indicating no discernible trend towards AI. This lack of a clear trend made it impossible to analyze investor behavior regarding a trend-based bias linked to AI.

Furthermore, the temporal analysis using Kernel Density Estimates (KDE) revealed that VR was once a "hot" topic in the EdTech market. Interestingly, the association with this "hot" topic influenced investor behavior negatively. Ventures founded during the hype phase with growing interest in VR received less funding on average than both VR businesses across the entire dataset and the wider EdTech peer group founded within the same timeframe.

These findings challenge the notion of AI adoption as a dominant driver in venture creation within the EdTech market and highlight the importance of understanding the nuanced role of emerging technologies in shaping industry trends and investor behavior.

7 REFERENCES

Abrahamson, E., & Hambrick, D. C. (1997). Attentional homogeneity in industries: The effect of discretion. *Journal of Organizational Behavior*, *18*(S1), 513–532. https://doi.org/10.1002/(SICI)1099-1379(199711)18:1+<513::AID-JOB905>3.0.CO;2-8

Alti, A., & Tetlock, P. C. (2014). Biased Beliefs, Asset Prices, and Investment: A Structural Approach. *The Journal of Finance*, *69*(1), 325–361. https://doi.org/10.1111/jofi.12089

Barr, A. (2015, July 1). Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. *Wall Street Journal*. http://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

Berg, A., Buffie, E. F., & Zanna, L.-F. (2018, May). Should We Fear the Robot Revolution? (The Correct Answer is Yes). *IMF Working Paper*. https://www.imf.org/en/Publications/WP/Issues/2018/05/21/Should-We-Fear-the-Robot-Revolution-The-Correct-Answer-is-Yes-44923

Berndt, T., Froese, H., Leverkus, L., & Ornik, R. (2014). *Comply or Explain als Lösung? : Bewertungsunterschiede und Intransparenz analysiert - Verbesserungen identifiziert*. https://www.alexandria.unisg.ch/handle/20.500.14171/86438

Bessen, J., Impink, S. M., Reichensperger, L., & Seamans, R. (2022). The role of data for AI startup growth. *Research Policy*, *51*(5), 104513. https://doi.org/10.1016/j.respol.2022.104513

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1st ed). O'Reilly.

Blei, D. M., Ng, A., & Michael Jordan. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brynjolfsson, E., & McAfee, A. (2017, July 18). The Business of Artificial Intelligence. *Harvard Business Review*. https://hbr.org/2017/07/the-business-of-artificial-intelligence

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Buck, A., Herrmann, C., & Lubkowitz, D. (1998). *Handbuch Trendmanagement: Innovation und Ästhetik als Grundlage unternehmerischer Erfolge* (1. Aufl). Frankfurter Allgemeine Zeitung, Verl.-Bereich Buch.

Bureau of Labor Statistics. (2023, April). *New business success rate U.S. 2022*. Statista. https://www.statista.com/statistics/725044/survival-rate-new-business-united-states/

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), Article 1. https://doi.org/10.1038/s42003-022-03036-1

Cavalry Ventures. (2022, October 12). Valuation Model: Backsolving valuations for SaaS Companies at Series A (and beyond). *Cavalry Chronicle*. https://medium.com/cavalry-chronicle/valuation-model-backsolving-valuations-for-saas-companies-at-series-a-and-beyond-326155984fd1

CB Insights. (2023, October 19). *The Complete List Of Unicorn Companies*. https://instapage.cbinsights.com/research-unicorn-companies

Chae, B., & Olson, D. L. (2021). Discovering Latent Topics of Digital Technologies From Venture Activities Using Structural Topic Modeling. *IEEE Transactions on Computational Social Systems*, *8*(6), 1438–1449. https://doi.org/10.1109/TCSS.2021.3085715

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. http://arxiv.org/abs/2307.09009

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT Goes to Law School. *Journal of Legal Education*, *387*(2022). https://dx.doi.org/10.2139/ssrn.4335905

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018). *Notes from the AI Frontier Insights from Hundreds of Use Cases*. McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.pdf

Cole, D. (2023). The Chinese Room Argument. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/chinese-room/

Cooper, M. J., Dimitrov, O., & Rau, P. R. (2001). A Rose.com by Any Other Name. *The Journal of Finance*, *56*(6), 2371–2388. https://www.jstor.org/stable/2697826

Cowgill, B. (2019). *Bias and Productivity in Humans and Machines*. W.E. Upjohn Institute. https://doi.org/10.17848/wp19-309

Criscuolo, C., Gal, P. N., & Carlo Menon. (2014). The Dynamics of Employment Growth: New Evidence from 18 Countries. *OECD Science, Technology and Industry Policy Papers*, *14*(14). http://dx.doi.org/10.1787/5jz417hj6hg6-en

Cronqvist, H., & Siegel, S. (2014). The genetics of investment biases. *Journal of Financial Economics*, *113*(2), 215–234. https://doi.org/10.1016/j.jfineco.2014.04.004

Dalle, J.-M., Besten, M. den, & Menon, C. (2017). *Using Crunchbase for economic and managerial research*. OECD. https://doi.org/10.1787/6c418d60-en

Dennett, D. C. (2016). Illusionism as the Obvious Default Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11–12), 65–72.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Feng, C. (2022a, March 1). China tech crackdown: Beijing's off-campus tutoring ban puts 90 per cent of firms out of business. *South China Morning Post*.

Feng, C. (2022b, May 11). Beijing bans preschool learning apps as crackdown on private tutoring continues. *South China Morning Post*.

Fields, R., & Pan, A. (Directors). (2022, December 22). *Jerome Pesenti—Large Language Models, PyTorch, and Meta*. Weights & Biases. https://wandb.ai/wandb_fc/gradient-dissent/reports/Jerome-Pesenti-Large-Language-Models-PyTorch-and-Meta--VmlldzozMTgzMjc2?galleryTag=podcast

Gallagher, S., & Palmer, J. (2020, September 29). The Pandemic Pushed Universities Online. The Change Was Long Overdue. *Harvard Business Review*. https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue

Gartner. (2023, August 23). *4 Exciting New Trends in the Gartner Emerging Technologies Hype Cycle*. Gartner. https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies

Gassmann, O., Csik, M., & Frankenberger, K. (2014). *The Business Model Navigator: 55 Models That Will Revolutionise Your Business* (1st edition). FT Press.

Gassmann, O., Frankenberger, K., & Sauer, R. (2016). *Exploring the Field of Business Model Innovation: New Theoretical Perspectives*. Springer.

Gassmann, O., Haefner, N., & Wagner, D. (2023). *Universities in an Age of Uncertainty 44 Propositions on the Future of Universities*. https://www.alexandria.unisg.ch/handle/20.500.14171/117981

gluebenchmark.com. (2023). *GLUE Benchmark*. https://gluebenchmark.com/

Goff, P. (2020). *Galileo's error: Foundations for a new science of consciousness* (First Vintage books edition). Vintage Books.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Google. (2023, September 28). *Google Trends*. Explore Search Interest for Artificial Intelligence by Time, Location and Popularity on Google Trends. https://trends.google.com/trends/explore?date=today%205-y&q=%2Fm%2F0mkz

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Guzman, J., & Li, A. (2023). Measuring Founding Strategy. *Management Science*, *69*(1), 101–118. https://doi.org/10.1287/mnsc.2022.4369

Gyeonggi Northern Chamber of Commerce and Industry. (2021, March 2). *South Korea: Survival rate of businesses compared to OECD by age*. Statista.

https://www.statista.com/statistics/1323672/south-korea-survival-rate-of-businesses-compared-to-oecd-by-age/

Haefner, N., Parida, V., Gassmann, O., & Wincent, J. (2023). Implementing and scaling artificial intelligence: A review, framework, and research agenda. *Technological Forecasting and Social Change*, *197*, 122878. https://doi.org/10.1016/j.techfore.2023.122878

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*(1), 5–28. https://doi.org/10.1007/BF00057884

Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, *124*(5), 1423–1465. https://doi.org/10.1086/688176

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

ISO. (n.d.). *Online Browsing Platform*. Retrieved August 15, 2023, from https://www.iso.org/obp/ui/#search

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or Not? A Gamified Approach to the Turing Test* (arXiv:2305.20010). arXiv. https://doi.org/10.48550/arXiv.2305.20010

Kahneman, D., & Riepe, M. W. (1998). Aspects of Investor Psychology. *The Journal of Portfolio Management*, *24*(4), 52–65. https://doi.org/10.3905/jpm.1998.409643

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Koller, T., Goedhart, M., & Wessels, D. (2005). The Right Role for Multiples in Valuation. *McKinsey on Finance*, *15*(Spring 2005), 7–11. https://papers.ssrn.com/abstract=805166

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Liang, Y. E., & Yuan, S.-T. D. (2016). Predicting investor funding behavior using crunchbase social network features. *Internet Research*, *26*(1), 74–100. https://doi.org/10.1108/IntR-09-2014-0231

Lyons, R. K. (2017). Economics of the Ed Tech Revolution. *California Management Review*, *59*(4), 49–55. https://doi.org/10.1177/0008125617717708

Marche, S. (2022, December 6). The College Essay Is Dead. *The Atlantic*. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

Mason, H., Stewart, D., & Gill, B. (1958, November 28). Rival. *The New Yorker*. https://www.newyorker.com/magazine/1958/12/06/rival-2

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. https://web.archive.org/web/20230210114240/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Meta. (2021, October 28). The Facebook Company Is Now Meta. *Meta*. https://about.fb.com/news/2021/10/facebook-company-is-now-meta/

Michel, A. H. (2023, August 16). Inside the messy ethics of making war with machines. *MIT Technology Review*. https://www.technologyreview.com/2023/08/16/1077386/war-machines/

Miric, M., Jia, N., & Huang, K. G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, *44*(2), 491–519. https://doi.org/10.1002/smj.3441

Mitchell, J., & Bowers, J. (2020). Priorless Recurrent Networks Learn Curiously. *Proceedings of the 28th International Conference on Computational Linguistics*, 5147–5158. https://doi.org/10.18653/v1/2020.coling-main.451

Moore, G. E. (1998). Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE*, *86*(1), 82–85. https://doi.org/10.1109/JPROC.1998.658762

OECD. (2017). *Entrepreneurship at a Glance 2017*. http://dx.doi.org/10.1787/entrepreneur_aag-2017-en

Oppy, G., & Dowe, D. (2021). The Turing Test. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entriesuring-test/

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pomerleau, D. A. (1988). ALVINN: An Autonomous Land Vehicle in a Neural Network. *Advances in Neural Information Processing Systems, 1*. https://proceedings.neurips.cc/paper/1988/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html

Prahalad, C. K., & Bettis, R. A. (1986). The Dominant Logic: A New Linkage between Diversity and Performance. *Strategic Management Journal*, *7*(6), 485–501. https://doi.org/10.1002/smj.4250070602

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed). Wiley-Interscience.

Retterath, A., & Braun, R. (2020). *Benchmarking Venture Capital Databases* (SSRN Scholarly Paper 3706108). https://doi.org/10.2139/ssrn.3706108

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386–408. https://doi.org/10.1037/h0042519

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Russakovsky, O., Jia Deng, Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sahlmann, W., & Scherlis, D. (2009). *Method for Valuing High-Risk, Long-Term Investments: The "Venture Capital Method."* Harvard Business School Background Note 288-006.

Schisgall, E. J., & Hamid, R. D. (2023, June 20). CS50 Will Integrate Artificial Intelligence Into Course Instruction | News. *The Harvard Crimson*. https://www.thecrimson.com/article/2023/6/21/cs50-artificial-intelligence/

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610. https://doi.org/10.1093/biomet/66.3.605

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning* (arXiv:2202.05924). arXiv. https://doi.org/10.48550/arXiv.2202.05924

Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. *MIS Quarterly*, *40*(4), 1035–1056. https://www.jstor.org/stable/26629687

Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, *17*(1), 83–104. https://doi.org/10.1257/089533003321164967

Spärck Jones, K. (1972). *A Statistical Interpretation of Term Specificity and its Application in Retrieval*. https://doi.org/10.1108/eb026526

Stevenson, H. H., Roberts, M. J., & Bhide, A. (1999). *The Entrepreneurial Venture* (W. A. Sahlman, Ed.; Second edition). Harvard Business Review Press.

Terwiesch, C. (2023). Would Chat GPT3 Get a Wharton MBA? *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*.

Thompson, A. D. (2021, July 9). Timeline of AI and language models. *Life Architect*. https://lifearchitect.ai/timeline/

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://www.jstor.org/stable/2251299

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

UBS. (2023). *Let's chat about ChatGPT*. https://www.ubs.com/us/en/wealth-management/insights/market-news/article.1585717.html

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Waldrop, M. M. (2016). More than Moore. *Nature News*, *530*(7589), 144–147. https://doi.org/10.1038/530144a

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language

Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/10.18653/v1/W18-5446

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Whitford, E. (2023, January 18). ChatGPT And AI Will Fuel New EdTech Boom. *Forbes*. https://www.forbes.com/sites/emmawhitford/2023/01/18/chatgpt-and-ai-will-fuel-new-edtech-boom/

WIPO. (2022). *Global Innovation Index 2022 results*. WIPO. https://www.wipo.int/pressroom/en/articles/2022/article_0011.html

Zeng, X., Li, Y., Leung, S. C. H., Lin, Z., & Liu, X. (2016). Investment behavior prediction in heterogeneous information network. *Neurocomputing*, *217*, 125–132. https://doi.org/10.1016/j.neucom.2015.12.139

Zhong, H., Liu, C., Zhong, J., & Xiong, H. (2018). Which startup to invest in: A personalized portfolio strategy. *Annals of Operations Research*, *263*(1), 339–360. https://doi.org/10.1007/s10479-016-2316-z

DISCLOSURE OF DIGITAL TOOLS

Note: wherever possible the connected academic source was cited (e.g. Bird, Loper, and Klein (2009) were cited in APA7 format in chpt. 3.2.2 instead of only specifying "nltk" in the following table)

tool :	**in-text location**	**usage:**	**cited:**
ChatGPT **(GPT-3.5)**	Full text	Generate draft based on key words: Prompt schema: "Draft a paragraph, ranging between 50 and 120 words in length, for an academic paper based on the following key insights: ****"	NO
Grammarly	Full text	Correcting spelling mistakes, suggest academic verbiage	NO
Pandas	3 Methodology, 4 Results	Data handling	NO
NumPy	3 Methodology, 4 Results	Mathematical functionality	NO

matplotlib	3 Methodology, 4 Results,	Illustration	YES
seaborn	3 Methodology, 4 Results	Illustration	YES
google.colab	3 Methodology, 4 Results	Google cloud resources for topic modeling	NO
nltk	3 Methodology, 4 Results	Language processing	YES
BERTopic	3 Methodology, 4 Results	Topic modeling	YES
sklearn	3 Methodology, 4 Results	Machine learning	YES

DECLARATION OF AUTHORSHIP

I hereby declare,

• that I have written this thesis independently,

• that I have written the thesis using only the aids specified in the index;

• that all parts of the thesis produced with the help of aids have been precisely declared;

• that I have mentioned all sources used and cited them correctly according to established academic citation rules;

• that I have acquired all immaterial rights to any materials I may have used, such as images or graphics, or that these materials were created by me;

• that the topic, the thesis, or parts of it have not already been the object of any work or examination of another course unless this has been expressly agreed with the faculty member in advance and is stated as such in the thesis;

• that I am aware of the legal provisions regarding the publication and dissemination of parts or the entire thesis and that I comply with them accordingly;

• that I am aware that my thesis can be electronically checked for plagiarism and for third-party authorship of human or technical origin and that I hereby grant the University of St.Gallen the copyright according to the Examination Regulations as far as it is necessary for the administrative actions;

• that I am aware that the University will prosecute a violation of this Declaration of Authorship and that disciplinary as well as criminal consequences may result, which may lead to expulsion from the University or to the withdrawal of my title."

By submitting this thesis, I confirm through my conclusive action that I am submitting the Declaration of Authorship, that I have read and understood it, and that it is true.

Mario Mühlematter

APPENDIX: NOTEBOOK PRINT-OUTS

Note: The complete code used in this thesis is also available under the following link: https://github.com/MarioMuehlematter/Education.ai

The appended printed files are not as easy to reuse, and the PDF export generates multiple formatting issues, including missing page numbers.

2022-04-22_Joining data

Joining individual data sets¶

The data sets in question are:

- 2022-04-22_EdTechs (2009, Europe)
- 2022-04-22_EdTechs (2009, US)
- 2022-04-22_EdTechs (2010, Europe)
- 2022-04-22_EdTechs (2010, US after Jan 1)
- 2022-04-22_EdTechs (2010, US)
- 2022-04-22_EdTechs (2011, Europe)
- 2022-04-22_EdTechs (2011, US)
- 2022-04-22_EdTechs (2011, US after Jan 1)
- 2022-04-22_EdTechs (2012, Europe)
- 2022-04-22_EdTechs (2012, US)
- 2022-04-22_EdTechs (2012, US after Jan 1)
- 2022-04-22_EdTechs (2013, Europe)
- 2022-04-22_EdTechs (2013, US)
- 2022-04-22_EdTechs (2013, US after Jan 1)
- 2022-04-22_EdTechs (2014, Europe)
- 2022-04-22_EdTechs (2014, US)
- 2022-04-22_EdTechs (2014, US after Jan 1)
- 2022-04-22_EdTechs (2015, Europe)
- 2022-04-22_EdTechs (2015, US)
- 2022-04-22_EdTechs (2015, US after Jan 1)
- 2022-04-22_EdTechs (2016, Europe)
- 2022-04-22_EdTechs (2016, US)
- 2022-04-22_EdTechs (2017, Europe)
- 2022-04-22_EdTechs (2017, US)
- 2022-04-22_EdTechs (2018, Europe)
- 2022-04-22_EdTechs (2018, US)
- 2022-04-22_EdTechs (2019)
- 2022-04-22_EdTechs (2020)
- 2022-04-22_EdTechs (2021-2022)
- 2022-04-25_EdTechs (2009, Asia)
- 2022-04-25_EdTechs (2010, Asia)
- 2022-04-25_EdTechs (2011, Asia)
- 2022-04-25_EdTechs (2012, Asia)

- 2022-04-25_EdTechs (2013, Asia)
- 2022-04-25_EdTechs (2014, Asia)
- 2022-04-25_EdTechs (2015 Feb-Dec, Asia)
- 2022-04-25_EdTechs (2015 Jan, Asia)
- 2022-04-25_EdTechs (2016, Asia)
- 2022-04-25_EdTechs (2017, Asia)
- 2022-04-25_EdTechs (2018, Asia)
- 2022-04-25_EdTechs (2019, Asia)
- 2022-04-25_EdTechs (2020-2022, Asia)

We're going to concatenate all the individual data sets together to create one master data set.

First let's import dependencies:

In [1]:

```
import pandas as pd
```

Let's create an array with all the data set names so we can loop over them and load them all:

In [2]:

```
datasets = ['2022-04-22_EdTechs (2009, Europe)', '2022-04-22_EdTechs (2009, US)', '2022-04-22_EdTechs (2010, Europe)', '2022-04-22_EdTechs (2010, US after Jan 1)', '2022-04-22_EdTechs (2010, US)', '2022-04-22_EdTechs (2011, Europe)', '2022-04-22_EdTechs (2011, US)', '2022-04-22_EdTechs (2011, US after Jan 1)', '2022-04-22_EdTechs (2012, Europe)', '2022-04-22_EdTechs (2012, US)', '2022-04-22_EdTechs (2012, US after Jan 1)', '2022-04-22_EdTechs (2013, Europe)', '2022-04-22_EdTechs (2013, US)', '2022-04-22_EdTechs (2013, US after Jan 1)', '2022-04-22_EdTechs (2014, Europe)', '2022-04-22_EdTechs (2014, US)', '2022-04-22_EdTechs (2014, US after Jan 1)', '2022-04-22_EdTechs (2015, Europe)', '2022-04-22_EdTechs (2015, US)', '2022-04-22_EdTechs (2015, US after Jan 1)', '2022-04-22_EdTechs (2016, Europe)', '2022-04-22_EdTechs (2016, US)', '2022-04-22_EdTechs (2017, Europe)', '2022-04-22_EdTechs (2017, US)', '2022-04-22_EdTechs (2018, Europe)', '2022-04-22_EdTechs (2018, US)', '2022-04-22_EdTechs (2019)', '2022-04-22_EdTechs (2020)', '2022-04-22_EdTechs (2021-2022)', '2022-04-25_EdTechs (2009, Asia)', '2022-04-25_EdTechs (2010, Asia)', '2022-04-25_EdTechs (2011, Asia)', '2022-04-25_EdTechs (2012, Asia)', '2022-04-25_EdTechs (2013, Asia)', '2022-04-25_EdTechs (2014, Asia)', '2022-04-25_EdTechs (2015 Feb-Dec, Asia)', '2022-04-25_EdTechs (2015 Jan, Asia)', '2022-04-25_EdTechs (2016, Asia)', '2022-04-25_EdTechs (2017, Asia)', '2022-04-25_EdTechs (2018, Asia)', '2022-04-25_EdTechs (2019, Asia)', '2022-04-25_EdTechs (2020-2022, Asia)']
```

In [3]:

```
print(len(datasets))
```

```
42
```

Now we create an array to store all our loaded datasets:

In [4]:

```
df_arr = []

for dataset in datasets:
    df_arr.append(pd.read_csv(dataset + '.csv'))
```

In [5]:

```
print(len(df_arr))
```

```
42
```

Joining the dataframes together:

In [6]:

```
df = pd.concat(df_arr, ignore_index=True)
```

In [7]:

```
print(df.shape)
```

```
(28997, 113)
```

Saving the dataframe:

In [8]:

```
df.to_csv('2022-04-25_EdTech data set.csv', index=False)
```

In [ ]:

Education.ai_Pre-Processing

Dataset Pre-Processing

Before starting the NLP-based analysis of the data, some general data-cleaning is in order.

Importing Libraries and Loading the Data¶

In [ ]:

```
#import necessary lib
import pandas as pd
```

In [ ]:

```
#load dataset
from google.colab import drive
drive.mount('1LbVccBPxo8pR9C41wosD5qYOglKsDrPr')
```

```
Mounted at 1LbVccBPxo8pR9C41wosD5qYOglKsDrPr
```

In [ ]:

```
df =
pd.read_csv("/content/1LbVccBPxo8pR9C41wosD5qYOglKsDrPr/MyDrive/001_Ma
sterarbeit/2022-12-31_EdTech data set_raw.csv", index_col=0)
```

```
<ipython-input-4-c0ebe79655f4>:1: DtypeWarning: Columns
(76,77,79,82,103,106,111) have mixed types. Specify dtype option on
import or set low_memory=False.
  df =
pd.read_csv("/content/1LbVccBPxo8pR9C41wosD5qYOglKsDrPr/MyDrive/001_Ma
sterarbeit/2022-12-31_EdTech data set_raw.csv", index_col=0)
```

First Look¶

In [ ]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 29346 entries, D'Alessandris to The Refined-U Digital Academy
Columns: 112 entries, Organization Name URL to Aberdeen - IT Spend
Currency (in USD)
dtypes: float64(34), object(78)
memory usage: 25.3+ MB
```

In [ ]:

```
for i in df.columns: print(i)
```

```
Organization Name URL
Founded Date
Founded Date Precision
Industries
Headquarters Location
Description
CB Rank (Company)
Estimated Revenue Range
Operating Status
Exit Date
Exit Date Precision
Closed Date
Closed Date Precision
Company Type
Website
Twitter
LinkedIn
Full Description
Facebook
Headquarters Regions
Investment Stage
Investor Type
Number of Portfolio Organizations
Number of Investments
Number of Lead Investments
Number of Diversity Investments
Number of Exits
Number of Exits (IPO)
Accelerator Program Type
Accelerator Application Deadline
Accelerator Duration (in weeks)
School Type
School Program
Number of Enrollments
School Method
Number of Founders (Alumni)
Industry Groups
Number of Founders
Founders
Number of Employees
Number of Funding Rounds
Funding Status
Last Funding Date
Last Funding Amount
Last Funding Amount Currency
Last Funding Amount Currency (in USD)
Last Funding Type
Last Equity Funding Amount
Last Equity Funding Amount Currency
Last Equity Funding Amount Currency (in USD)
Last Equity Funding Type
Total Equity Funding Amount
Total Equity Funding Amount Currency
Total Equity Funding Amount Currency (in USD)
Total Funding Amount
Total Funding Amount Currency
Total Funding Amount Currency (in USD)
Top 5 Investors
Number of Lead Investors
Number of Investors
Number of Acquisitions
Acquisition Status
Transaction Name
Transaction Name URL
Acquired by
Acquired by URL
Announced Date
Announced Date Precision
Price
Price Currency
Price Currency (in USD)
Acquisition Type
Acquisition Terms
IPO Status
IPO Date
Delisted Date
Delisted Date Precision
Money Raised at IPO
Money Raised at IPO Currency
Money Raised at IPO Currency (in USD)
Valuation at IPO
Valuation at IPO Currency
Valuation at IPO Currency (in USD)
Stock Symbol
Stock Symbol URL
Stock Exchange
Last Leadership Hiring Date
Last Layoff Mention Date
Number of Events
SEMrush - Monthly Visits
SEMrush - Average Visits (6 months)
SEMrush - Monthly Visits Growth
SEMrush - Visit Duration
SEMrush - Visit Duration Growth
SEMrush - Page Views / Visit
SEMrush - Page Views / Visit Growth
SEMrush - Bounce Rate
SEMrush - Bounce Rate Growth
SEMrush - Global Traffic Rank
SEMrush - Monthly Rank Change (#)
SEMrush - Monthly Rank Growth
BuiltWith - Active Tech Count
Apptopia - Number of Apps
Apptopia - Downloads Last 30 Days
G2 Stack - Total Products Active
IPqwery - Patents Granted
IPqwery - Trademarks Registered
IPqwery - Most Popular Patent Class
IPqwery - Most Popular Trademark Class
Aberdeen - IT Spend
Aberdeen - IT Spend Currency
Aberdeen - IT Spend Currency (in USD)
```

As the list above indicates there are two columns labelled description (decription/ full description). Too keep as much information as possible filtering for NaN is run on both and the column with less invalid values is kept for further processing.

In [ ]:

```
#convert descriptions to lowercase
df['Full Description'].str.lower()
df['Description'].str.lower()
```

Out[ ]:

```
Organization Name
D'Alessandris                    d'alessandris is a group of companies
that off...
Berg Marin AB                    berg marin is a technology
information company...
NoTosh                           we help people think differently, so
they can ...
The CI Bookshop                  the ci bookshop offers easy readers
and teachi...
Navigationsgruppen               the company will train professional
and recrea...
                                                       ...

Careerboat                       careerboat is a career acceleration
platform f...
UpVisit
paas
Finstep                                   mobile app, financial
education for genz
HubChannel                       uma plataforma de monetização através
de conte...
The Refined-U Digital Academy    edtech solutions, networking and
```

```
educational s...
Name: Description, Length: 29346, dtype: object
```

In [ ]:

```
#remove NaN values in description/ full description column and
establish temporary dataframes for comparison
temp1 = df[df['Full Description'].notnull()]
temp2 = df[df['Description'].notnull()]
```

In [ ]:

```
temp1['Word Count'] = temp1['Full Description'].str.split().str.len()
temp2['Word Count'] = temp2['Description'].str.split().str.len()
```

```
<ipython-input-9-b2ec766b316c>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  temp1['Word Count'] = temp1['Full
Description'].str.split().str.len()
<ipython-input-9-b2ec766b316c>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  temp2['Word Count'] = temp2['Description'].str.split().str.len()
```

Filtering the Descriptions¶

Excluded companies whose descriptions are too short (word counts <30) to be useful for topic modeling, following the example of Chae & Olson (2021)

In [ ]:

```
temp1 = temp1[temp1['Word Count'] >= 30]
temp2 = temp2[temp2['Word Count'] >= 30]
```

In [ ]:

```
#essetial summary statistics for the cleand datasets

print('The dataset temp1 has ' + str(temp1.shape[0]) + ' entries.\n'+
      'The descriptions in the dataset contain a total of '+
str(temp1['Word Count'].sum()) +' words.')
print('The dataset temp2 has ' + str(temp2.shape[0]) + ' entries.\n'+
      'The descriptions in the dataset contain a total of '+
str(temp2['Word Count'].sum()) +' words.')
```

```
The dataset temp1 has 14046 entries.
The descriptions in the dataset contain a total of 1225060 words.
The dataset temp2 has 0 entries.
The descriptions in the dataset contain a total of 0 words.
```

All entries in the simple description column are too short, according to established data science practice. Therfore the 'Full Description' cloumn will be used for further NLP processing.

In [ ]:

```
temp1['Word Count'].mean()
```

Out[ ]:

```
87.21771322796526
```

In [ ]:

```
temp1['Founded Date']. min()
```

Out[ ]:

```
'01.01.09'
```

In [ ]:

```
#export cleaned dataset
from google.colab import files
temp1.to_csv('2023-12-31_EdTech dataset_cleaned.csv')
files.download('2023-12-31_EdTech dataset_cleaned.csv')
```

193d691028c548098375df8274fcfc4b

Analysing Company Descriptions using NLP

```
#import libs
import pandas as pd
import numpy as np
import nltk
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
```

```
True
```

```
#mount drive
from google.colab import drive
drive.mount('1LbVccBPxo8pR9C41wosD5qYOglKsDrPr')
```

```
Mounted at 1LbVccBPxo8pR9C41wosD5qYOglKsDrPr
```

```
df =
pd.read_csv("/content/1LbVccBPxo8pR9C41wosD5qYOglKsDrPr/MyDrive/001_Ma
sterarbeit/2022-12-31_EdTech dataset_cleaned.csv", index_col=0)
```

```
<ipython-input-4-ebc2da93ffb4>:1: DtypeWarning: Columns (77,103,106)
have mixed types. Specify dtype option on import or set
low_memory=False.
  df =
pd.read_csv("/content/1LbVccBPxo8pR9C41wosD5qYOglKsDrPr/MyDrive/001_Ma
sterarbeit/2022-12-31_EdTech dataset_cleaned.csv", index_col=0)
```

```
#add numeric index
df = df.reset_index()
```

```
#add the 'ID' column to the original df to enable merging later on
df['ID']=df.index
```

#Topic Modelling with BERTopic

```
pip install bertopic
```

```
Collecting bertopic
  Downloading bertopic-0.15.0-py2.py3-none-any.whl (143 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.4/143.4 kB 2.0 MB/s eta
0:00:00
```

```
ent already satisfied: numpy>=1.20.0 in
/usr/local/lib/python3.10/dist-packages (from bertopic) (1.23.5)
Collecting hdbscan>=0.8.29 (from bertopic)
  Downloading hdbscan-0.8.33.tar.gz (5.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 5.2/5.2 MB 12.5 MB/s eta
0:00:00
ents to build wheel ... etadata (pyproject.toml) ... ap-learn>=0.5.0
(from bertopic)
  Downloading umap-learn-0.5.3.tar.gz (88 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 88.2/88.2 kB 6.8 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: pandas>=1.1.5 in
/usr/local/lib/python3.10/dist-packages (from bertopic) (1.5.3)
Requirement already satisfied: scikit-learn>=0.22.2.post1 in
/usr/local/lib/python3.10/dist-packages (from bertopic) (1.2.2)
Requirement already satisfied: tqdm>=4.41.1 in
/usr/local/lib/python3.10/dist-packages (from bertopic) (4.66.1)
Collecting sentence-transformers>=0.4.1 (from bertopic)
  Downloading sentence-transformers-2.2.2.tar.gz (85 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 86.0/86.0 kB 7.3 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: plotly>=4.7.0 in
/usr/local/lib/python3.10/dist-packages (from bertopic) (5.15.0)
Requirement already satisfied: cython<3,>=0.27 in
/usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29-
>bertopic) (0.29.36)
Requirement already satisfied: scipy>=1.0 in
/usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29-
>bertopic) (1.10.1)
Requirement already satisfied: joblib>=1.0 in
/usr/local/lib/python3.10/dist-packages (from hdbscan>=0.8.29-
>bertopic) (1.3.2)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas>=1.1.5->bertopic)
(2023.3)
Requirement already satisfied: tenacity>=6.2.0 in
/usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic)
(8.2.3)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from plotly>=4.7.0->bertopic)
(23.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-
learn>=0.22.2.post1->bertopic) (3.2.0)
Collecting transformers<5.0.0,>=4.6.0 (from sentence-
transformers>=0.4.1->bertopic)
  Downloading transformers-4.32.0-py3-none-any.whl (7.5 MB)
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.5/7.5 MB 31.7 MB/s eta
0:00:00
ent already satisfied: torch>=1.6.0 in /usr/local/lib/python3.10/dist-
packages (from sentence-transformers>=0.4.1->bertopic) (2.0.1+cu118)
Requirement already satisfied: torchvision in
/usr/local/lib/python3.10/dist-packages (from sentence-
transformers>=0.4.1->bertopic) (0.15.2+cu118)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-
packages (from sentence-transformers>=0.4.1->bertopic) (3.8.1)
Collecting sentencepiece (from sentence-transformers>=0.4.1->bertopic)
  Downloading sentencepiece-0.1.99-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 40.5 MB/s eta
0:00:00
 sentence-transformers>=0.4.1->bertopic)
  Downloading huggingface_hub-0.16.4-py3-none-any.whl (268 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 268.8/268.8 kB 19.6 MB/s eta
0:00:00
ent already satisfied: numba>=0.49 in /usr/local/lib/python3.10/dist-
packages (from umap-learn>=0.5.0->bertopic) (0.56.4)
Collecting pynndescent>=0.5 (from umap-learn>=0.5.0->bertopic)
  Downloading pynndescent-0.5.10.tar.gz (1.1 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.1/1.1 MB 36.9 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0-
>sentence-transformers>=0.4.1->bertopic) (3.12.2)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0-
>sentence-transformers>=0.4.1->bertopic) (2023.6.0)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0-
>sentence-transformers>=0.4.1->bertopic) (2.31.0)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0-
>sentence-transformers>=0.4.1->bertopic) (6.0.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.10/dist-packages (from huggingface-hub>=0.4.0-
>sentence-transformers>=0.4.1->bertopic) (4.7.1)
Requirement already satisfied: llvmlite<0.40,>=0.39.0dev0 in
/usr/local/lib/python3.10/dist-packages (from numba>=0.49->umap-
learn>=0.5.0->bertopic) (0.39.1)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from numba>=0.49->umap-
learn>=0.5.0->bertopic) (67.7.2)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1-
>pandas>=1.1.5->bertopic) (1.16.0)
Requirement already satisfied: sympy in
/usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-
transformers>=0.4.1->bertopic) (1.12)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-
transformers>=0.4.1->bertopic) (3.1)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-
transformers>=0.4.1->bertopic) (3.1.2)
Requirement already satisfied: triton==2.0.0 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.6.0->sentence-
transformers>=0.4.1->bertopic) (2.0.0)
Requirement already satisfied: cmake in
/usr/local/lib/python3.10/dist-packages (from triton==2.0.0-
>torch>=1.6.0->sentence-transformers>=0.4.1->bertopic) (3.27.2)
Requirement already satisfied: lit in /usr/local/lib/python3.10/dist-
packages (from triton==2.0.0->torch>=1.6.0->sentence-
transformers>=0.4.1->bertopic) (16.0.6)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from
transformers<5.0.0,>=4.6.0->sentence-transformers>=0.4.1->bertopic)
(2023.6.3)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1 (from
transformers<5.0.0,>=4.6.0->sentence-transformers>=0.4.1->bertopic)
  Downloading tokenizers-0.13.3-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (7.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.8/7.8 MB 34.3 MB/s eta
0:00:00
 transformers<5.0.0,>=4.6.0->sentence-transformers>=0.4.1->bertopic)
  Downloading safetensors-0.3.2-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 54.5 MB/s eta
0:00:00
ent already satisfied: click in /usr/local/lib/python3.10/dist-
packages (from nltk->sentence-transformers>=0.4.1->bertopic) (8.1.7)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
/usr/local/lib/python3.10/dist-packages (from torchvision->sentence-
transformers>=0.4.1->bertopic) (9.4.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.6.0-
>sentence-transformers>=0.4.1->bertopic) (2.1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.4.0->sentence-transformers>=0.4.1->bertopic) (3.2.0)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.4.0->sentence-transformers>=0.4.1->bertopic) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.4.0->sentence-transformers>=0.4.1->bertopic) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->huggingface-
hub>=0.4.0->sentence-transformers>=0.4.1->bertopic) (2023.7.22)
Requirement already satisfied: mpmath>=0.19 in
/usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.6.0-
>sentence-transformers>=0.4.1->bertopic) (1.3.0)
Building wheels for collected packages: hdbscan, sentence-
transformers, umap-learn, pynndescent
  Building wheel for hdbscan (pyproject.toml) ... e=hdbscan-0.8.33-
cp310-cp310-linux_x86_64.whl size=3039188
sha256=cd02187e72916a3be550e8e96993db192f745ee620b27a86eef072d455e93a6
6
  Stored in directory:
/root/.cache/pip/wheels/75/0b/3b/dc4f60b7cc455efaefb62883a7483e76f09d0
6ca81cf87d610
  Building wheel for sentence-transformers (setup.py) ... ers:
filename=sentence_transformers-2.2.2-py3-none-any.whl size=125924
sha256=26c2a21389365cc76c3342cfd2c87a3bddaae88107b14c6aa0af792230b5b87
3
  Stored in directory:
/root/.cache/pip/wheels/62/f2/10/1e606fd5f02395388f74e7462910fe851042f
97238cbbd902f
  Building wheel for umap-learn (setup.py) ... ap-learn:
filename=umap_learn-0.5.3-py3-none-any.whl size=82808
sha256=3300cf735ae9be16fde7eeee90045ad1dd90622ce1b9aa1c0b20de1af58d93e
5
  Stored in directory:
/root/.cache/pip/wheels/a0/e8/c6/a37ea663620bd5200ea1ba0907ab3c217042c
1d035ef606acc
  Building wheel for pynndescent (setup.py) ... e=pynndescent-0.5.10-
py3-none-any.whl size=55617
sha256=d6e6087e98db425a3885e7f94a8e79b011bdf4f10d8f3ad76494df17a6622df
3
  Stored in directory:
/root/.cache/pip/wheels/4a/38/5d/f60a40a66a9512b7e5e83517ebc2d1b42d857
be97d135f1096
Successfully built hdbscan sentence-transformers umap-learn
pynndescent
Installing collected packages: tokenizers, sentencepiece, safetensors,
huggingface-hub, transformers, pynndescent, hdbscan, umap-learn,
sentence-transformers, bertopic
Successfully installed bertopic-0.15.0 hdbscan-0.8.33 huggingface-hub-
0.16.4 pynndescent-0.5.10 safetensors-0.3.2 sentence-transformers-
2.2.2 sentencepiece-0.1.99 tokenizers-0.13.3 transformers-4.32.0 umap-
learn-0.5.3
```

```python
from bertopic import BERTopic
from sklearn.feature_extraction.text import CountVectorizer

vectorizer_model = CountVectorizer(stop_words="english")
min_cluster_size = 100
```

```
#define model parameters
topic_model = BERTopic(vectorizer_model=vectorizer_model,
min_topic_size = min_cluster_size)

#initiate model
docs = df['Full Description']
topics, probs = topic_model.fit_transform(docs)
```

```
{"model_id":"cb28400e04e143cc8daa5a9742c5b8a1","version_major":2,"vers
ion_minor":0}

{"model_id":"2c252435572941b897aaeec1f33a28f2","version_major":2,"vers
ion_minor":0}

{"model_id":"dbd2bf49e1264eea93057eecf9f305e1","version_major":2,"vers
ion_minor":0}

{"model_id":"8aa07e43d6f340b48968ba2d13a82f0c","version_major":2,"vers
ion_minor":0}

{"model_id":"4efaf9d324c4476898296ac2f545584d","version_major":2,"vers
ion_minor":0}

{"model_id":"8091343c11eb49a79f9efb66f244833b","version_major":2,"vers
ion_minor":0}

{"model_id":"4b086eb0fa5b4d22b7fef7a2e036c96b","version_major":2,"vers
ion_minor":0}

{"model_id":"1c05a662fbf743c0b976cb0ebe8477f6","version_major":2,"vers
ion_minor":0}

{"model_id":"1147750a65ef45c292f90d26f919fe4c","version_major":2,"vers
ion_minor":0}

{"model_id":"3031bd185192493eac36af5e4e8b21ea","version_major":2,"vers
ion_minor":0}

{"model_id":"bf34b6fbda6a436096368d0446b8166d","version_major":2,"vers
ion_minor":0}

{"model_id":"75cbf9a2168946089b4305087f6b2a15","version_major":2,"vers
ion_minor":0}

{"model_id":"919aea8f80b440758d9fdec44d58a09b","version_major":2,"vers
ion_minor":0}

{"model_id":"6e4ba647842c4ae0bb706fea92796c58","version_major":2,"vers
ion_minor":0}
```

```
#extract results as a df
df_results = pd.DataFrame(topic_model.get_document_info(docs))
```

```
#add results to cleaned dataset
df = df.join(df_results)

df[df['Topic']>=0]
```

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
2	Olyssa	https://www.crunch	07.12.09	day	Education, Human Res	Athens, Attiki, Greec	Educational jobfindi	1,057,344	NaN	Active	...	135	2	There is an obvious ga	6	6_career_job_talen	[career, job, talent,	[FindMyProfessioni	career-job-talent-	1.000000	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
	base.com/organization/			ources,Internet	e	ngnetwork							pofcommunicationbetwe..		t_candidates	candidates,jobs,employ	sanelitecareerservice...	candidates-jobs-em...		

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
		olyssa												.			...				
3	Vaibmu	https://www.crun	01.12.09	day	Classifieds,Consu	Espoo,SouthernFi	Vaibmuisacompany	1,615,632	NaN	Active	...	336	3	Vaibmuisastrateg	3	3_services_solut	[services,soluti	[Digileapisconfide	services-solutio	0.868778	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
	chbase.com/organizatio			lting, Training	nland, Finland	that provides management							ic consulting & portfolio		ions_development_busin	ons, development, business	nt of making its customers	ns-development-business		

	OrganizationName	OrganizationNameURLn/vaibmu	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Descriptionc...	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Documentd...	Topic	Nameess	Representations,s...	Representative_Docs...	Top_n_wordss...	Probability	Representative_document
4	Low PriceLesso	https://www.cor	30.11.09	day	Automotive, E-C	Edinburgh, Edin	DrivingLessonc	737,771	Less than $1M	Closed	...	30	4	Low PriceLesso	8	8_books_readin	[books, reading	[Sellbuybook.co	books-reading-	0.722258	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
ns.com	unchbase.com/organizat			ommerce, Education	burgh, City of, United King	omparison website							ns.com is an on-line servi		g_book_readers	, book, readers, read, aut	m is an Online Portal where	book-readers-read-auth		

	Organization Name	Organization Name URL	Founded Date	Founded Date Precision	Industries	Headquarters Location	Description	CB Rank (Company)	Estimated Revenue Range	Operating Status	...	Word Count	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
		ion/low-pr...				dom								ceth...			hors,...	peo...	...		
7	LocusAssig	https://ww	12.11.09	day	Consulting	Coventry,C	HND Assign	467,170	NaN	Active	...	48	7	LocusAssig	15	15_writing	[writing,w	[GetAssign	writing-wr	0.720963	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
nments	w.crunchbase.com/organ			,Education,ServiceIndu	oventry,UnitedKingdom	mentsUK is now LocusAss							nmentsistopclassassign		_writers_essay_assignm	riters,essay,assignmen	mentHelponAnyTopicbyth	iters-essay-assignment		

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
		ization/hnd-as...			stry		ignmentspr...							mentwrit...		ent	t, academic...	eirExp...	-acade...		
8	Gradua	https:	05.11.	day	Educat	Copenh	Gradua	60,305	$1M to	Active	...	218	8	Gradua	6	6_care	[caree	[FindM	career	0.7678	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
teland	//www.crunchbase.com/o	09		ion,Recruiting,SocialM	agen,Hovedstaden,Denma	telandoffersacareernet		$10M					telandisEurope'slargest		er_job_talent_candidat	r,job,talent,candidates	yProfessionisanelitecar	-job-talent-candidates	34	

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank(Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Names	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
		rganization/gradua...			edia	rk	workforstude...							tcareernetwor...		es	,jobs,employ...	eeserservice...	-jobs-em...		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

	Organization Name	Organization Name URL	Founded Date	Founded Date Precision	Industries	Headquarters Location	Description	CB Rank (Company)	Estimated Revenue Range	Operating Status	...	Word Count	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
14037	Pronounce	https://www.crunchbase	01.05.22	month	Artificial Intelligence, E	NaN	Pronounce is a personalize	978,772	NaN	Active	...	36	14037	Pronounce is a personalize	4	4_language_english_lear	[language, english, learn	[SkimaTalk is an online lan	language-english-learni	0.363318	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
	.com/organization/pron			-Learning, Education...		dspeechanalyzerth...							dspeechanalyzerth...		ning_languages	ing, languages, speak...	guage-learningplat...	ng-languages-sp...		

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
		ounce																			
14038	SkywinITAcademy	https://www.crunc	01.05.22	day	Education	Surat, Gujarat, India	SkywinITAcademy is	1,483,675	NaN	Closed	...	39	14038	SkywinITTraining,	19	19_training_training_sap_s	[training,sap,sof	[IntroductiontoII	training-sap-soft	1.000000	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
	hbase.com/organization					highereducationacademy							Suratbasedoneofthelea..		oftware_courses	tware,courses,institut	CTIICTisaleadingTraini.	ware-courses-institut...		

	Organization Name	Organization Name URL/skywin...	Founded Date	Founded Date Precision	Industries	Headquarters Location	Description...	CB Rank (Company)	Estimated Revenue Range	Operating Status	...	Word Count	ID	Document.	Topic	Name	Representation,...	Representative_Docs..	Top_n_words	Probability	Representative_document
14041	KidFi	https://www.c	28.04.22	day	Apps, Educatio	Baku, Baki, Azer	KidFi teachesk	154,001	NaN	Active	...	53	14041	A gamified mobil	0	0_children_kid	[children, kids	[Gofamilingisa	children-kids-	0.742120	False

Organization Name	Organization Name URL	Founded Date	Founded Date Precision	Industries	Headquarters Location	Description	CB Rank (Company)	Estimated Revenue Range	Operating Status	...	Word Count	ID	Document	Topic	Names	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
	runchbase.com/organiza			n, Gamification	baijan	ids financial literacy and h							e application is designed f		s_parents_games	, parents, games, child, le	Start-Up that is born with	parents-games-child-le..		

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	DocumentTopic	Topic	Name	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
		tion/kidfi					elp...							or...			arni...	th...	.		
14043	Firstep	https://www.c	27.04.22	day	EdTech, Educat	Hamburg, Hambu	Mobile App, Fina	341,346	NaN	Active	...	214	14043	Firstephelpsy	11	11_financial_l	[financial, loan	[Webelievefinan	financial-loan	1.00000	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
	runchbase.com/organiza			ion, FinTech	rg, Germany	ncial Education for GenZ							oung people between the ag		oan_money_loans	, money, loans, student, cr	cial advice and financial p	-money-loans-student-c		

	OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CBRank(Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Documents	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
		tion/finstep												esof...			edi...	la...	...		
14045	TheRefined-U	https://www	25.04.22	day	EdTech,Educ	Abuja,Feder	Edtechsolut	1,656,060	NaN	Active	...	33	14045	Onlineeducat	10	10_educatio	[educations,	[EdSurgeist	education-s	0.698795	False

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank(Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representations	Representative_Docs	Top_n_words	Probability	Representative_document
DigitalAcademy	.crunchbase.com/organi			ation,Training	alCapitalTerritory,Nig	ions,networkingandeduc							ionalsupportserviceage		n_students_online_educa	students,online,educati	heleadingsiteforeducat	tudents-online-educatio		

OrganizationName	OrganizationNameURL	FoundedDate	FoundedDatePrecision	Industries	HeadquartersLocation	Description	CB Rank (Company)	EstimatedRevenueRange	OperatingStatus	...	WordCount	ID	Document	Topic	Name	Representation	Representative_Docs	Top_n_words	Probability	Representative_document
	zation/the-re...				eria	ationals...							ncy for...		tional	onal, pla...	ors, en...	nal-...		

5694 rows × 123 columns

#Bag-of-Word method using 'scikit-learn Random Forest'

The goal is to represent each document (patent) based on the number of times each word occurs in them. We are going to use the Scikit learn (https://scikit-learn.org/stable/index.html) package, which has built-in functions that make this process very easy and straightforward.

This method has been proven to be the best performing in classifing patents mentioning AI by Miric, M., Jia, N. and Huang, K. (Forthcoming) Using Supervised Machine Learning for Large-Scale Classification In Management Research: The Case For Identifying Artificial Intelligence Patents. Strategic Management Journal.

Link: https://www.milanmiric.com/ai-patents-us

Further Reading:

- short explanation of BOW
- https://thatascience.com/learn-machine-learning/bag-of-words/
- short explanation of RandomForest classifier
- https://thatascience.com/learn-machine-learning/build-random-forest/

Convert Text Data to Term-Document Frequency Matrix:

We are going to use the built in sklearn tools to process the text. The idea is to construct a matrix represetantion of the data. The rows represent each document (patent abstract). The columns represent each word that occurs in any of the documents (corpus). The fields of the matrix (TF) represent how often each word occurs in each document.

This is illustrated below.

TFIDF.png

```
## import libs

# Sklearn Packages #
from sklearn.feature_extraction.text import TfidfTransformer,
CountVectorizer, TfidfVectorizer
from sklearn.metrics import mean_squared_error, r2_score,
accuracy_score, classification_report
from sklearn.model_selection import train_test_split,
StratifiedShuffleSplit, StratifiedKFold, cross_val_predict
from sklearn.metrics import accuracy_score, roc_auc_score,
precision_score, recall_score, f1_score, confusion_matrix

# Import SkLearn Classifiers #
from sklearn.ensemble import RandomForestClassifier

nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
True
```

```python
#define the text pre-processing
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from gensim.utils import simple_preprocess
import re


lemmatizer = WordNetLemmatizer()

cachedStopWords = set(stopwords.words('english'))
def pre_process(text):
    words = simple_preprocess(text,deacc=True,min_len=3)
    text = [lemmatizer.lemmatize(re.sub('[^A-Za-z]+', ' ', word)) for
word in words if not word in cachedStopWords]
    text = ' '.join(map(str,text))
    return text

#apply pre-process to text of 'Full Description
df['BOW'] = df.apply(lambda row : pre_process(row['Full
Description']), axis = 1)

#additional stopwords based on previous model outputs and update
stopwords list
#new_stopwords = ['']
#cachedStopWords = pd.concat(cachedStopWords,new_stopwords)

# Define Model Parameters #
# Minimum Document Frequency -- Minimum number of times a word needs
to occur to be considered #
MINDF = 10

# Maximum Document Frequency -- Maximum share of documents where a
word needs to occur to be considered #
MAXDF = 0.8

# Maximum number of features we would want to consider -- ranked by
most frequently occuring #
MF=1200

# NGrams -- Number of Word Pairs. Takes the form (Min, Max). E.g. (1,
2) means single words and word pairs #
NGrams = (1,2)

# Stopwords -- List of common words we want to omit (STOP_WORDS) #
from spacy.lang.en import STOP_WORDS
STOP_WORDS = list(STOP_WORDS)

# Define Tokenizer -- This is a Custom Function that we use #
```

```
from textblob import TextBlob

# Define Textblob Tokenizer -- This Converts Words to Lowercase and
Stems the keywords #
def textblob_tokenizer(str_input):
    blob = TextBlob(str_input.lower())
    tokens = blob.words
    words = [token.stem() for token in tokens]
    return words

# Define Vectorizer #
vec = TfidfVectorizer(max_features= MF,
                      max_df = MAXDF,
                      stop_words = STOP_WORDS,
                      ngram_range=(1, 2),
                      tokenizer=textblob_tokenizer)
```

In the following paragraph the training data (from Miric et al.) is used to train the model before applying it to the EdTech dataset. To clearly mark variables relating to training, they are marked '_t' at the end.

```
# Load Training Data from Maric et al (2023)
#available at: xxxx
df_t =
pd.read_csv("/content/1LbVccBPxo8pR9C41wosD5qYOglKsDrPr/MyDrive/001_Ma
sterarbeit/Copy of 4K Patents - AI 20p.csv")

#appl pre-process to text
df_t['BOW'] = df_t.apply(lambda row : pre_process(row['abstract']),
axis = 1)

# Perform vectorizatiion and extract feature names on the training
data
txt_t = list(df_t['BOW'])
txt_vec_t = vec.fit_transform(txt_t)
FEATURENAMES_t = vec.get_feature_names_out()

#create a Bag of Words df showing vectorized values and tokens of the
training dataset
df_bow_t = pd.DataFrame(txt_vec_t.toarray(), columns = FEATURENAMES_t)

#add actual classification to the df
df_bow_t['CLASS'] = df_t['actual']

df_bow_t.sample()
```

	absorb	abut	acceler	access	accommod	accord	accordembodi	accordingli	account	accuraci	...	window	wire	wireless	work	workpiec	wrap	write	zone	zoom	CLASS
2656	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

1 rows × 1201 columns

```
# Prepare training data for building the model
X_train = df_bow_t.drop(['CLASS'], axis=1)
y_train = df_bow_t['CLASS']

# Instantiate the model
rfc = RandomForestClassifier(n_estimators= 1000)

# Train/Fit the model
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier(n_estimators=1000)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

RandomForestClassifier

```
RandomForestClassifier(n_estimators=1000)
```

After training the model on the pre-existing data set from Miric et al, the model can be applied to the EdTech data set.

```
#apply vectorizer to Full Description column of EdTech data
txt = list(df['BOW'])
```

```
txt_vec = vec.fit_transform(txt)
FEATURENAMES = vec.get_feature_names_out()

#create a Bag of Words df showing vectorized values and tokens
df_bow = pd.DataFrame(txt_vec.toarray(), columns = FEATURENAMES)

#generate a classification dataframe
count = 0
df_cls = pd.DataFrame([],columns = ['Class'])

while count < len(df_bow):
  df_temp = pd.DataFrame([])
  X_pred = df_bow.iloc[count]
  y_pred = rfc.predict([X_pred])
  df_temp['Class'] = y_pred
  df_cls = df_cls.append(df_temp)
  count = count +1

df_cls = df_cls.reset_index(drop = True)

df_cls
```

	Class
0	0
1	0
2	0
3	1
4	0
...	...
14041	0
14042	0
14043	0
14044	0
14045	0

14046 rows × 1 columns

```
#merege the topics table with the inital dataset
df = df.join(df_cls, how='left')

#exporting the dataset before further analysis
from google.colab import files
```

```
df.to_csv('2022-12-31_EdTech dataset_processed.csv')
files.download('2022-12-31_EdTech dataset_processed.csv')
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.Javascript object>
```

Education.ai_Analysis+Plots

Import and Basic Settings¶

In [ ]:

```
import os as os
os.getcwd()
```

Out[ ]:

```
'/Users/mario/Documents/Github/Education.ai/Education.ai/notebooks'
```

Analyse the metrics described in the paper

- Total funding
- No. of Employees
- Description
- Classification (incl. self-declared)
- Age [extract date - funding date]

In [ ]:

```
#import libs
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:

```
'''
#mount google drive
from google.colab import drive
drive.mount('1LbVccBPxo8pR9C41wosD5qYOglKsDrPr')
'''
```

Out[ ]:

```
"\n#mount google drive\nfrom google.colab import drive\
ndrive.mount('1LbVccBPxo8pR9C41wosD5qYOglKsDrPr')\n"
```

In [ ]:

```
#import processed dataset
df =
pd.read_csv("//Users/mario/Documents/Github/Education.ai/data/2022-12-
31_EdTech dataset_processed.csv", index_col=0)
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/2151554003.py:2: DtypeWarning: Columns (78,104,107)
have mixed types. Specify dtype option on import or set
low_memory=False.
  df =
pd.read_csv("//Users/mario/Documents/Github/Education.ai/data/2022-12-
31_EdTech dataset_processed.csv", index_col=0)
```

In [ ]:

```
#basic setting for plot layout
sns.set_theme(style= 'darkgrid' )#"whitegrid")
plt.rcParams.update({'font.size': 22})
my_colors = 'Set1' #'flag' #'brg'   #'BrBG_r' # "ch:r=-.2,d=.3_r"
```

In [ ]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 14046 entries, 0 to 14045
Columns: 125 entries, Organization Name to Class
dtypes: bool(1), float64(36), int64(4), object(84)
memory usage: 13.4+ MB
```

In [ ]:

```
for i in df.columns: print(i)
```

```
Organization Name
Organization Name URL
Founded Date
Founded Date Precision
Industries
Headquarters Location
Description
CB Rank (Company)
Estimated Revenue Range
Operating Status
Exit Date
Exit Date Precision
Closed Date
Closed Date Precision
Company Type
Website
Twitter
LinkedIn
Full Description
Facebook
Headquarters Regions
Investment Stage
Investor Type
```

```
Number of Portfolio Organizations
Number of Investments
Number of Lead Investments
Number of Diversity Investments
Number of Exits
Number of Exits (IPO)
Accelerator Program Type
Accelerator Application Deadline
Accelerator Duration (in weeks)
School Type
School Program
Number of Enrollments
School Method
Number of Founders (Alumni)
Industry Groups
Number of Founders
Founders
Number of Employees
Number of Funding Rounds
Funding Status
Last Funding Date
Last Funding Amount
Last Funding Amount Currency
Last Funding Amount Currency (in USD)
Last Funding Type
Last Equity Funding Amount
Last Equity Funding Amount Currency
Last Equity Funding Amount Currency (in USD)
Last Equity Funding Type
Total Equity Funding Amount
Total Equity Funding Amount Currency
Total Equity Funding Amount Currency (in USD)
Total Funding Amount
Total Funding Amount Currency
Total Funding Amount Currency (in USD)
Top 5 Investors
Number of Lead Investors
Number of Investors
Number of Acquisitions
Acquisition Status
Transaction Name
Transaction Name URL
Acquired by
Acquired by URL
Announced Date
Announced Date Precision
Price
Price Currency
Price Currency (in USD)
Acquisition Type
Acquisition Terms
IPO Status
IPO Date
Delisted Date
Delisted Date Precision
Money Raised at IPO
Money Raised at IPO Currency
Money Raised at IPO Currency (in USD)
Valuation at IPO
Valuation at IPO Currency
Valuation at IPO Currency (in USD)
Stock Symbol
Stock Symbol URL
Stock Exchange
Last Leadership Hiring Date
Last Layoff Mention Date
Number of Events
SEMrush - Monthly Visits
SEMrush - Average Visits (6 months)
SEMrush - Monthly Visits Growth
SEMrush - Visit Duration
SEMrush - Visit Duration Growth
SEMrush - Page Views / Visit
SEMrush - Page Views / Visit Growth
SEMrush - Bounce Rate
SEMrush - Bounce Rate Growth
SEMrush - Global Traffic Rank
SEMrush - Monthly Rank Change (#)
SEMrush - Monthly Rank Growth
BuiltWith - Active Tech Count
Apptopia - Number of Apps
Apptopia - Downloads Last 30 Days
G2 Stack - Total Products Active
IPqwery - Patents Granted
IPqwery - Trademarks Registered
IPqwery - Most Popular Patent Class
IPqwery - Most Popular Trademark Class
Aberdeen - IT Spend
Aberdeen - IT Spend Currency
Aberdeen - IT Spend Currency (in USD)
Word Count
ID
Document
Topic
Name
Representation
Representative_Docs
Top_n_words
Probability
Representative_document
```

```
BOW
Class
```

In [ ]:

```
df['Total Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
16354647.550672183
```

Formating Data for Analysis¶

In [ ]:

```
#rename columns to be more descriptive
df = df.rename(columns={'Name': 'Topic Name'})
```

In [ ]:

```
#defining age calculation
from datetime import date

cut_off = date(2022, 12, 31)

def calculateAge(x):
    t0 = cut_off
    age = t0.year - x.year -((t0.month, t0.day) <
         (x.month, x.day))

    return age
```

In [ ]:

```
#define a age calculation with a continous value
def calculateAgeDays(x):
  t0 = cut_off
  age = sum([(t0.year- x.year) * 365.25, (t0.month - x.month) * 30,
(t0.day - x.day)])

  return age
```

In [ ]:

```
#converting str in 'Founded Date' to datetime format
from pandas.core.reshape.tile import to_datetime
df['Founded Date'] = to_datetime(df['Founded Date'])
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/1797541815.py:3: UserWarning: Could not infer format,
so each element will be parsed individually, falling back to
```

```
`dateutil`. To ensure parsing is consistent and as-expected, please
specify a format.
  df['Founded Date'] = to_datetime(df['Founded Date'])
```

In [ ]:

```
#generating 'Age' col from 'Founded Date' col
df['Age in Years'] = df.apply(lambda row : calculateAge(row['Founded
Date']), axis = 1)
```

In [ ]:

```
#generating 'Age in Days' col from 'Founded Date' col
df['Age in Days'] = df.apply(lambda row :
calculateAgeDays(row['Founded Date']), axis = 1)
```

In [ ]:

```
#converting classification outcome to boolean valuables
df['Class'] = df['Class'].astype(bool)
```

In [ ]:

```
#split the location cell into actionable parts
df[['City','Country']]  = df['Headquarters Location'].str.rsplit(r",",
n=1, expand=True)
```

In [ ]:

```
#The 'Industries' field offers companies/ data curators the option to
specify AI in a profile, extract boolean value for said field
df['Class self-identified'] =
df['Industries'].str.contains('Artificial Intelligence', regex= False)
```

In [ ]:

```
#define a function to turn numbers wrongfully identified as dates back
into the classification buckets
def emp_corr(var):
  i = str(var)
  if i == '01.Oct':
    return '1-10'
  elif i =='Nov.50':
    return '11-50'
  else:
    return var
```

In [ ]:

```
#apply the above function to the 'Number of Employees' column
df['Number of Employees'] = df.apply(lambda row : emp_corr(row['Number
of Employees']), axis = 1)
```

Origin Country of Ventures¶

Check if findings of Dalle et al. (2017) hold true for the dataset

- mostly US-based *(how is it worded exactlly? majority or just biggest part?)*

In [ ]:

```
#count the values for Countries
countries = df['Country'].value_counts()
```

In [ ]:

```
#print top 10 countries of origin (overview as appendix)
countries[:10]
```

Out[ ]:

```
Country
 United States      6405
 India              1951
 United Kingdom     1231
 France              357
 Germany             342
 Spain               327
 China               314
 Israel              218
 Switzerland         192
 The Netherlands     190
Name: count, dtype: int64
```

In [ ]:

```
#reshape 'countries Series' for better visibility in the plot
#the top 10
countries_plt= countries[:10].copy()

#append 'others' category
others = pd.Series({' Others' : countries[10:].sum()})
countries_plt = pd.concat([countries_plt,(others)])

#add column names
countries_plt = pd.DataFrame(countries_plt, copy = True, columns = ['#
Ventures'])
countries_plt = countries_plt.reset_index(names='Country')
```

In [ ]:

```
#insert a column with the ISO 3166 Alpha-3 code for country names to
increase readability
countries_plt['Origin'] = ['USA', 'IND', 'GBR', 'FRA','DEU', 'ESP',
'CHN','ISR', 'CHE', 'NLD','others']
```

In [ ]:

```
#calculate the sample size
countries_plt['# Ventures'].sum()
```

Out[ ]:

```
14004
```

In [ ]:

```
# bar plot
sns.barplot(x = 'Origin' , y = '# Ventures', data = countries_plt,
palette= my_colors,)
plt.show()
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/1470902161.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(x = 'Origin' , y = '# Ventures', data = countries_plt,
palette= my_colors,)
```

No description has been provided for this image

Note: the assumption found in theory seems to hold true

Size of Ventures¶

- mostly small (<50 FTE) firms in sample.

In [ ]:

```
#count the values for small firms
size = df['Number of Employees'].value_counts()
```

In [ ]:

```
size
```

Out[ ]:

```
Number of Employees
1-10         6987
11-50        4619
51-100        720
101-250       483
251-500       217
501-1000      103
```

```
1001-5000       71
10001+          32
5001-10000      16
Name: count, dtype: int64
```

In [ ]:

```
#calculate the sample size
size.sum()
```

Out[ ]:

```
13248
```

In [ ]:

```
#calculate percentage of ventures with less than 50 employees
(size['1-10']+ size['11-50'])/size.sum()
```

Out[ ]:

```
0.8760567632850241
```

Survival Rate of Ventures¶

In [ ]:

```
#Import .xlsx with Data from Statista
df_sr =
pd.read_excel("/Users/mario/Documents/Github/Education.ai/data/surviva
l-rates.xlsx",
                    sheet_name = 0)
```

In [ ]:

```
df_sr['Sample Count'] = df['Age in Years'].value_counts()
```

In [ ]:

```
survivors = df[df['Operating Status']== 'Active']
df_sr['Sample Survivors'] = survivors['Age in Years'].value_counts()
```

In [ ]:

```
#calculate the survival rate
df_sr['Sample'] = (df_sr['Sample Survivors'] / df_sr['Sample Count'])
* 100
```

In [ ]:

```
sr_plt = sns.lineplot(data = df_sr[['OECD','USA','Sample']],
                      palette = my_colors,
                      linewidth = 2,
                      marker='o',
                      linestyle= '--')
sr_plt.set_ylim(0, 100)
sr_plt.set_ylabel('Survival Rate in %')
sr_plt.set_xlabel('Age in Years')
```

Out[ ]:

```
Text(0.5, 0, 'Age in Years')
```

No description has been provided for this image

Note: OECD values are older; the US rates show the 'Corona-Freeze' in bankrupcy observed also in other economies

Furthermore, bias identified (survivor/ "publsihing bias") in Retterath & Braun 2020

Topic Model (Unsupervised Learning)¶

In [ ]:

```
df['Topic Name'].value_counts()
```

Out[ ]:

```
Topic Name
-1_learning_students_platform_education       8352
0_children_kids_parents_games                  698
1_entrepreneurs_startup_startups_business      589
2_college_students_universities_university     435
3_services_solutions_development_business      375
4_language_english_learning_languages          370
5_sports_fitness_athletes_coaches              332
6_career_job_talent_candidates                 258
7_coding_code_programming_learn                244
8_books_reading_book_readers                   213
9_reality_vr_virtual_ar                        210
10_education_students_online_educational       202
11_financial_loan_money_loans                  195
12_medical_healthcare_health_medicine          181
13_tutors_tutoring_tutor_students              171
14_video_users_content_videos                  168
15_writing_writers_essay_assignment            166
16_school_academy_education_schools            147
17_learning_teachers_students_data             143
18_education_chinese_china_beijing             137
19_training_sap_software_courses               130
20_test_exam_exams_prep                        120
21_music_piano_musical_guitar                  107
```

```
22_security_cyber_cybersecurity_services       103
Name: count, dtype: int64
```

In [ ]:

```
#top10 best funded ventures ("fund returners")
top10 = df['Total Funding Amount Currency (in
USD)'].nlargest(10).index.to_list()
top10_extract = df.loc[top10]
top10_sorted = top10_extract[['Organization Name', 'Topic Name','Total
Funding Amount Currency (in USD)']].sort_values('Total Funding Amount
Currency (in USD)',ascending= False)
```

In [ ]:

```
top10_sorted
```

Out[ ]:

	Organization Name	Topic Name	Total Funding Amount Currency (in USD)
10676	BYJU’S	-1_learning_students_platform_education	6.987658e+09
11007	Yuanfudao	18_education_chinese_china_beijing	4.052000e+09
11663	Zuoyebang	18_education_chinese_china_beijing	2.935000e+09
2354	CommonBond	11_financial_loan_money_loans	1.559106e+09
10544	Eruditus Executive Education	-1_learning_students_platform_education	1.163986e+09
11368	VIPKID	-1_learning_students_platform_education	1.055077e+09
12042	Emeritus	-1_learning_students_platform_ed	1.040000e+09

	Organization Name	Topic Name	Total Funding Amount Currency (in USD)
		ucation	
10546	Zhangmen	18_education_chinese_china_beijing	9.022773e+08
11317	DaDa	18_education_chinese_china_beijing	8.638673e+08
12044	Unacademy	-1_learning_students_platform_education	8.385000e+08

In [ ]:

```
sns.barplot(data = top10_sorted, x= 'Organization Name', y= 'Total
Funding Amount Currency (in USD)', hue= 'Topic Name',
palette=my_colors)
plt.xticks(rotation = 45, wrap = True, ha = 'right')
```

Out[ ]:

```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9],
 [Text(0, 0, 'BYJU’S'),
  Text(1, 0, 'Yuanfudao'),
  Text(2, 0, 'Zuoyebang'),
  Text(3, 0, 'CommonBond'),
  Text(4, 0, 'Eruditus Executive Education'),
  Text(5, 0, 'VIPKID'),
  Text(6, 0, 'Emeritus'),
  Text(7, 0, 'Zhangmen'),
  Text(8, 0, 'DaDa'),
  Text(9, 0, 'Unacademy')])
```

No description has been provided for this image

8352 entries could not be fitted into a categorie of more than 100 entries. Topic '-1' is thus dropped from the analysis. See BERTopic documentation.

In [ ]:

```
#Exclude Topic '-1' from analysis
topic_df = df[df['Topic']!= -1]
```

In [ ]:

```
x = topic_df[['Topic Name', 'Total Funding Amount Currency (in
USD)']].groupby('Topic Name').mean()
x = x.sort_values(by= ['Total Funding Amount Currency (in USD)'],
ascending= False)
print(x)
```

```
                                            Total Funding Amount
Currency (in USD)
Topic Name

18_education_chinese_china_beijing
1.868380e+08
11_financial_loan_money_loans
4.046202e+07
16_school_academy_education_schools
3.907554e+07
22_security_cyber_cybersecurity_services
1.898600e+07
7_coding_code_programming_learn
1.263014e+07
5_sports_fitness_athletes_coaches
1.233157e+07
17_learning_teachers_students_data
1.112621e+07
4_language_english_learning_languages
8.411668e+06
12_medical_healthcare_health_medicine
7.869882e+06
19_training_sap_software_courses
7.656393e+06
20_test_exam_exams_prep
7.617884e+06
9_reality_vr_virtual_ar
7.531751e+06
1_entrepreneurs_startup_startups_business
7.308803e+06
14_video_users_content_videos
6.956945e+06
10_education_students_online_educational
6.911315e+06
15_writing_writers_essay_assignment
6.394808e+06
21_music_piano_musical_guitar
5.992062e+06
6_career_job_talent_candidates
5.481909e+06
2_college_students_universities_university
5.268921e+06
0_children_kids_parents_games
4.892457e+06
```

```
3_services_solutions_development_business
4.599133e+06
8_books_reading_book_readers
4.164009e+06
13_tutors_tutoring_tutor_students
2.820270e+06
```

In [ ]:

```
sns.barplot(data = x, x= 'Topic Name', y= 'Total Funding Amount
Currency (in USD)', palette=my_colors)
plt.xticks(rotation = 45, wrap = True, ha = 'right')
plt.yscale('log')
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/1473611591.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(data = x, x= 'Topic Name', y= 'Total Funding Amount
Currency (in USD)', palette=my_colors)
```

No description has been provided for this image

In [ ]:

```
sns.despine(left=True, bottom=True)

sns.displot(
    data=topic_df,
    x='Founded Date' ,#'Age in Days',#"Age"
    hue="Topic Name",
    kind="kde", height=6,
    multiple="fill", clip=(0, None),
    palette= my_colors,
)
```

Out[ ]:

```
<seaborn.axisgrid.FacetGrid at 0x284c1a790>
```

```
<Figure size 640x480 with 0 Axes>
```

No description has been provided for this image

Topics worth investigating: Expansion in 10, 15, 17 Decreased interest in 8, 9 Complete drop-off in 16, 18, 19

Notes, possible Explanations:

8: Books -> market saturation, shrinking, unatractive market?

9: VR -> META effect? resp. market trend influencing META

10: online education -> Coronoa Boom

15: writing -> LLM integrations -> possible AI link

16: -> moves with 18, same explanation?

17: -> follows 10, learning moved online, better data availabillity

18: China banned education apps

19: Software training -> Market saturation?

Thus seems to rather follow macro-trends, than technological topics

In [ ]:

```
rep_doc = df[df['Representative_document']== True]
```

In [ ]:

```
#create an export with the relevant columns to add as an appendix, 
showing the comapny name, the original description, the assigned topic 
and representative words
topic_export = (rep_doc[['Organization Name','Full Description','Topic 
Name', 'Top_n_words']])
topic_export.to_csv('//Users/mario/Documents/Github/Education.ai/data/
topic_model_exports.csv')
```

subsampling T18/T16¶

In [ ]:

```
df_t18 = df[df['Topic Name'] == '18_education_chinese_china_beijing']
```

In [ ]:

```
df_t18.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 137 entries, 649 to 13496
Columns: 130 entries, Organization Name to Class self-identified
dtypes: bool(3), datetime64[ns](1), float64(37), int64(4), object(85)
memory usage: 137.4+ KB
```

In [ ]:

```
df_t18['Country'].value_counts()
```

Out[ ]:

```
Country
 China                   103
 United States            11
 United Kingdom            6
 Taiwan                    3
 France                    2
 South Korea               2
 Hong Kong                 2
 India                     2
 Ireland                   1
 Singapore                 1
 Vietnam                   1
 Japan                     1
 Israel                    1
 United Arab Emirates      1
Name: count, dtype: int64
```

In [ ]:

```
df_t16 = df[df['Topic Name'] == '16_school_academy_education_schools']
```

In [ ]:

```
df_t16.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 147 entries, 305 to 12552
Columns: 130 entries, Organization Name to Class self-identified
dtypes: bool(3), datetime64[ns](1), float64(37), int64(4), object(85)
memory usage: 147.4+ KB
```

In [ ]:

```
df_t16['Country'].value_counts()
```

Out[ ]:

```
Country
 United States      101
 United Kingdom      25
 India                4
 France               3
 Spain                2
 Malaysia             2
 Vietnam              2
 Singapore            2
 Germany              1
 Switzerland          1
 The Netherlands      1
 Saudi Arabia         1
 Philippines          1
 China                   1
Name: count, dtype: int64
```

The overlap is either coincidental or the ventures served china from outside the country.

Subsampling '15_writing_writers_essay_assignment' after 2020-12-31¶

given the insights above subsampling topic 15 seems promising, as it appears to be a proxy for LLMs

In [ ]:

```
sub_df = df[df['Founded Date'] > '2020-12-31']
```

In [ ]:

```
sub_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 511 entries, 10095 to 14045
Columns: 130 entries, Organization Name to Class self-identified
dtypes: bool(3), datetime64[ns](1), float64(37), int64(4), object(85)
memory usage: 512.5+ KB
```

In [ ]:

```
sub_df[['Topic Name','Class self-identified']].groupby('Topic
Name').value_counts()
```

Out[ ]:

```
Topic Name                                 Class self-identified
-1_learning_students_platform_education    False
302
                                           True
32
0_children_kids_parents_games              False
22
                                           True
1
10_education_students_online_educational   False
11
                                           True
1
11_financial_loan_money_loans              False
9
12_medical_healthcare_health_medicine      False
4
13_tutors_tutoring_tutor_students          False
```

```
7
14_video_users_content_videos               False
8
15_writing_writers_essay_assignment         False
9
17_learning_teachers_students_data          False
8
19_training_sap_software_courses            False
1
1_entrepreneurs_startup_startups_business   False
18
20_test_exam_exams_prep                     False
5
21_music_piano_musical_guitar               False
2
22_security_cyber_cybersecurity_services    False
4
2_college_students_universities_university  False
11
3_services_solutions_development_business   False
11
                                            True
1
4_language_english_learning_languages       False
7
                                            True
4
5_sports_fitness_athletes_coaches           False
7
6_career_job_talent_candidates              False
9
                                            True
1
7_coding_code_programming_learn             False
9
8_books_reading_book_readers                False
3
9_reality_vr_virtual_ar                     False
4
Name: count, dtype: int64
```

In [ ]:

```
sub_df[['Topic Name','Class']].groupby('Topic Name').value_counts()
```

Out[ ]:

```
Topic Name                                  Class
-1_learning_students_platform_education     False    312
                                            True      22
0_children_kids_parents_games               False     20
```

```
                                            True       3
10_education_students_online_educational    False     12
11_financial_loan_money_loans               False      7
                                            True       2
12_medical_healthcare_health_medicine       False      4
13_tutors_tutoring_tutor_students           False      6
                                            True       1
14_video_users_content_videos               False      6
                                            True       2
15_writing_writers_essay_assignment         False      7
                                            True       2
17_learning_teachers_students_data          False      8
19_training_sap_software_courses            False      1
1_entrepreneurs_startup_startups_business   False     16
                                            True       2
20_test_exam_exams_prep                     False      4
                                            True       1
21_music_piano_musical_guitar               False      2
22_security_cyber_cybersecurity_services    False      4
2_college_students_universities_university  False     11
3_services_solutions_development_business   False     12
4_language_english_learning_languages       False      9
                                            True       2
5_sports_fitness_athletes_coaches           False      6
                                            True       1
6_career_job_talent_candidates              False      9
                                            True       1
7_coding_code_programming_learn             False      9
8_books_reading_book_readers                False      2
                                            True       1
9_reality_vr_virtual_ar                     False      4
Name: count, dtype: int64
```

In [ ]:

```
a = sub_df[sub_df['Topic Name'] ==
'15_writing_writers_essay_assignment' ]
b = sub_df[sub_df['Topic Name'] !=
'15_writing_writers_essay_assignment' ]
```

In [ ]:

```
a.info()
print('-----------')
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 9 entries, 10137 to 14020
Columns: 130 entries, Organization Name to Class self-identified
dtypes: bool(3), datetime64[ns](1), float64(37), int64(4), object(85)
memory usage: 9.0+ KB
```

```
-----------
<class 'pandas.core.frame.DataFrame'>
Index: 502 entries, 10095 to 14045
Columns: 130 entries, Organization Name to Class self-identified
dtypes: bool(3), datetime64[ns](1), float64(37), int64(4), object(85)
memory usage: 503.5+ KB
```

In [ ]:

```
a['Total Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
24182.0
```

In [ ]:

```
b['Total Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
1191062.6952380952
```

In [ ]:

```
s =sub_df[['Topic Name', 'Total Funding Amount Currency (in
USD)']].groupby('Topic Name').mean()
s = s.sort_values(by= ['Total Funding Amount Currency (in USD)'],
ascending= False)
print(s)
```

```
                                            Total Funding Amount
Currency (in USD)
Topic Name

2_college_students_universities_university
3.446905e+06
6_career_job_talent_candidates
2.760858e+06
-1_learning_students_platform_education
1.432991e+06
1_entrepreneurs_startup_startups_business
9.403960e+05
11_financial_loan_money_loans
6.614013e+05
5_sports_fitness_athletes_coaches
6.047153e+05
7_coding_code_programming_learn
5.525870e+05
14_video_users_content_videos
3.515485e+05
3_services_solutions_development_business
```

```
3.203215e+05
0_children_kids_parents_games
1.953872e+05
4_language_english_learning_languages
1.750000e+05
9_reality_vr_virtual_ar
1.325000e+05
17_learning_teachers_students_data
5.000000e+04
15_writing_writers_essay_assignment
2.418200e+04
21_music_piano_musical_guitar
3.635000e+03
10_education_students_online_educational
NaN
12_medical_healthcare_health_medicine
NaN
13_tutors_tutoring_tutor_students
NaN
19_training_sap_software_courses
NaN
20_test_exam_exams_prep
NaN
22_security_cyber_cybersecurity_services
NaN
8_books_reading_book_readers
NaN
```

In [ ]:

```
sns.barplot(data = s, x= 'Topic Name', y= 'Total Funding Amount
Currency (in USD)', palette=my_colors)
plt.xticks(rotation = 45, wrap = True, ha = 'right')
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/1247354701.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(data = s, x= 'Topic Name', y= 'Total Funding Amount
Currency (in USD)', palette=my_colors)
```

Out[ ]:

```
([0,
  1,
  2,
  3,
  4,
  5,
  6,
  7,
  8,
  9,
  10,
  11,
  12,
  13,
  14,
  15,
  16,
  17,
  18,
  19,
  20,
  21],
 [Text(0, 0, '2_college_students_universities_university'),
  Text(1, 0, '6_career_job_talent_candidates'),
  Text(2, 0, '-1_learning_students_platform_education'),
  Text(3, 0, '1_entrepreneurs_startup_startups_business'),
  Text(4, 0, '11_financial_loan_money_loans'),
  Text(5, 0, '5_sports_fitness_athletes_coaches'),
  Text(6, 0, '7_coding_code_programming_learn'),
  Text(7, 0, '14_video_users_content_videos'),
  Text(8, 0, '3_services_solutions_development_business'),
  Text(9, 0, '0_children_kids_parents_games'),
  Text(10, 0, '4_language_english_learning_languages'),
  Text(11, 0, '9_reality_vr_virtual_ar'),
  Text(12, 0, '17_learning_teachers_students_data'),
  Text(13, 0, '15_writing_writers_essay_assignment'),
  Text(14, 0, '21_music_piano_musical_guitar'),
  Text(15, 0, '10_education_students_online_educational'),
  Text(16, 0, '12_medical_healthcare_health_medicine'),
  Text(17, 0, '13_tutors_tutoring_tutor_students'),
  Text(18, 0, '19_training_sap_software_courses'),
  Text(19, 0, '20_test_exam_exams_prep'),
  Text(20, 0, '22_security_cyber_cybersecurity_services'),
  Text(21, 0, '8_books_reading_book_readers')])
```

No description has been provided for this image

T9 VR¶

In [ ]:

```
df[df['Topic Name']== '9_reality_vr_virtual_ar']['Total Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
7531751.307692308
```

In [ ]:

```
#split data into hype-phase 2015-12-31 to 2019-12-31
df_hype = df[df['Founded Date'] > '2015-12-31']
df_hype = df_hype[df_hype['Founded Date']< '2019-12-31']
```

In [ ]:

```
#calculate the mean funding of VR during the hype phase
df_hype[df_hype['Topic Name']== '9_reality_vr_virtual_ar']['Total
Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
6572572.525
```

In [ ]:

```
#calculate the mean funding for all other topics during the hype phase
df_hype[df_hype['Topic Name']!= '9_reality_vr_virtual_ar']['Total
Funding Amount Currency (in USD)'].mean()
```

Out[ ]:

```
8564850.991485335
```

In [ ]:

```
#create df for illustration
hot_df = pd.DataFrame({'Category': ['all VR ventures', 'VR ventures
during hype-phase','other topics during hype-phase' ],
                       'Mean Total Funding': [7531751.307692308,
6572572.525,8564850.991485335]
})
```

In [ ]:

```
p = sns.barplot(
   data= hot_df,
   x= 'Category',
   y= 'Mean Total Funding',
   palette= my_colors)
plt.xticks(rotation = 45, wrap = True, ha = 'right')
p.set_xlabel('Subset')
```

```
/var/folders/sv/wh2gxdqd7l7g9jyshkbzflxr0000gn/T/
ipykernel_14444/3102664829.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
```

```
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  p = sns.barplot(
```

Out[ ]:

```
Text(0.5, 0, 'Subset')
```

No description has been provided for this image

Classification¶

In [ ]:

```
df['Class'].value_counts()
```

Out[ ]:

```
Class
False    12976
True      1070
Name: count, dtype: int64
```

In [ ]:

```
# Draw a scatter plot while assigning point colors and sizes to
different variables in the dataset
f, ax = plt.subplots(figsize=(16.5, 16.5))
sns.despine(f, left=True, bottom=True)
#ai_binary = {0: 'non-AI', 1: 'AI'}
size_ranking = ['10001+','5001-10000','1001-5000' ,'501-1000','251-
500','101-250','51-100','11-50','1-10']
sns.scatterplot(y="Total Funding Amount Currency (in USD)", x=
'Founded Date', #'Age in Days', #"Age"
                hue=  "Class",
                size="Number of Employees",
                palette= my_colors,
                #hue_order= ai_binary,
                size_order = size_ranking,
                sizes=(10, 400), linewidth=0,
                alpha=.7,
                data=df, ax=ax)
plt.yscale('log')
```

No description has been provided for this image

In [ ]:

```
df['Class self-identified'].value_counts()
```

Out[ ]:

```
Class self-identified
False    13550
True       496
Name: count, dtype: int64
```

In [ ]:

```
# Draw a scatter plot while assigning point colors and sizes to
different variables in the dataset
f, ax = plt.subplots(figsize=(16.5, 16.5))
sns.despine(f, left=True, bottom=True)
#ai_binary = {0: 'non-AI', 1: 'AI'}
size_ranking = ['10001+','5001-10000','1001-5000' ,'501-1000','251-
500','101-250','51-100','11-50','1-10']
sns.scatterplot(y="Total Funding Amount Currency (in USD)", x=
'Founded Date', #'Age in Days', #"Age"
                hue= 'Class self-identified' ,  #"Class"
                size="Number of Employees",
                palette= my_colors,
                #hue_order= ai_binary,
                size_order = size_ranking,
                sizes=(10, 400), linewidth=0,
                alpha=.7,
                data=df, ax=ax)
plt.yscale('log')
```

No description has been provided for this image

Notes:

1. Data Quality: a lot of entry dates coincide with the new years, quality is improving
2. AI companies see to be spread out randomly
3. as expected big, well-funded orgs are older (top left) -> seems valid

In [ ]:

```
#convert blue and red from 'Set11 to continous color scheme for
heatmap/ corr matrix
# Define the RGB codes as hex strings
color1 = "#e41a1c" #red
color2 = "#377eb8" #blue
```

In [ ]:

```
Set1_cmap = sns.light_palette(color= color2, as_cmap=True)
```

In [ ]:

```
#checking for corellation between analysed var and Total Funding
sns.heatmap(df[['Total Funding Amount Currency (in USD)', 'Class',
'Class self-identified',  'Topic']].corr(),
            annot= True,
            cmap = Set1_cmap
            )
plt.xticks(rotation = 45, wrap = True, ha = 'right')
plt.yticks(ha = 'right', wrap = True)
```

Out[ ]:

```
(array([0.5, 1.5, 2.5, 3.5]),
 [Text(0, 0.5, 'Total Funding Amount Currency (in USD)'),
  Text(0, 1.5, 'Class'),
  Text(0, 2.5, 'Class self-identified'),
  Text(0, 3.5, 'Topic')])
```

No description has been provided for this image

very low correlation for all analysed variables (Topic, AI, AI self-described)

These var do not correlate with successfull fundraising, the idea of Cooper et al. could not be replicated for the EdTech market