Understanding the Multidimensional Private Equity Space: Unsupervised Learning for Identifying Relationships in the Startup Ecosystem

Mark Rebotunov
twotensor.com

Stanislav Kharchenko
twotensor.com

Daniel Afshar
twotensor.com
daniel@twotensor.com

Anton Matskevich
twotensor.com

July 18, 2023

Abstract

Understanding the dynamics and relationships within the startup ecosystem is crucial for entrepreneurs, investors, policymakers, and researchers. In this research, we employ unsupervised learning techniques, specifically clustering and dimensionality reduction, to classify startup projects into meaningful domains. We assemble a comprehensive dataset of startup information from various sources, including Crunchbase, and fine-tune a T5 model to classify industry sectors based on text descriptions. With a dataset of approximately 100,000 observations, we explore dimensionality reduction techniques to identify the most informative features for clustering. Using the K-means algorithm, we determine the optimal number of clusters and apply it to the startup dataset. Our results reveal four distinct clusters, demonstrating meaningful separations among startups based on their attributes. The t-SNE visualization technique aids in understanding the relationships and patterns within each cluster. This research contributes to a deeper understanding of the startup ecosystem, empowering stakeholders to make informed decisions and foster the growth and success of startups.

1 Introduction

The dynamic evolution of the startup ecosystem continues to contribute significantly to global economic growth, permeating diverse industry sectors and

instigating an array of innovative solutions (Mason & Brown, 2014). Comprehending the multifaceted startup landscape, characterised by an array of complex and varying attributes, is paramount for a wide spectrum of stakeholders - entrepreneurs, investors, policymakers and academic researchers alike (Audretsch et al., 2016). Nevertheless, the intricate and constantly shifting entrepreneurial terrain imposes formidable challenges to meaningful analysis and domain classification.

Propelled by these challenges, the present study ventures to exploit the capabilities of machine learning techniques - specifically, clustering and dimensionality reduction - to tackle the complexities of startup domain classification. Clustering algorithms, representing a pillar of unsupervised learning, offer an efficient mechanism to group data instances based on their inherent features, circumventing the need for predetermined class labels (Jain, 2010). Concurrently, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) serve as critical tools for transforming high-dimensional data spaces into more manageable and interpretable formats without considerable loss of information (Van der Maaten & Hinton, 2008).

The central ambition of this investigation is to harness these computational techniques to facilitate the classification of startup projects into relevant domains. By drawing upon a comprehensive set of startup characteristics - encompassing elements like market orientation, technological deployment, team composition, geographical positioning, and business model - we aim to conceive an insightful domain classification framework. This framework is anticipated to unveil significant relationships and trends within the startup ecosystem that are not immediately discernible (Aghion et al., 2017).

The potential impact of this investigation spans multiple fronts. For entrepreneurs, comprehension of this domain classification can shed light on competitive dynamics, thereby informing strategic decisions (Blank, 2013). For investors, these insights can bolster decision-making processes and refine risk appraisal (Mollick, 2014). Policymakers can benefit from a more detailed grasp of the startup landscape to guide regulatory actions, while academics can use these findings as a stepping stone to expand the theoretical frontier of entrepreneurship (Zacharakis et al., 2003).

Through the amalgamation of clustering and dimensionality reduction techniques, this study endeavours to illuminate the complex tapestry of startup projects. Our approach is designed to offer a more refined, data-driven understanding of the dynamics within entrepreneurial activity, thereby enriching the extant body of knowledge concerning startup ecosystems.

2 Background

Startups are the lifeblood of the global economy, driving industry shifts and technological progress through their innovative solutions and disruptive models (Mason & Brown, 2014). The process of startup selection, nurturing, and fund-

ing is crucial to the entire entrepreneurial ecosystem, with venture capitalists (VCs) playing a pivotal role (Gompers et al., 2020). The task of understanding the multifaceted universe of startup projects is intricate, due to the wide array of characteristics these entities exhibit. This diversity presents unique challenges and opportunities for extracting deeper relationships and meaningfully categorizing startups, hence the necessity for advanced analytical frameworks (Audretsch et al., 2016).

Traditional startup classification, typically based on sectors, growth stages, and geographic location, has shown limited effectiveness due to the dynamic nature and complexity of the startup landscape (Blank, 2013). These classic decision-making paradigms often mirror those employed by VCs, who usually base their decisions on criteria such as the quality of the entrepreneurs and management team, the size and attractiveness of the market, and the product or technology of the firm (Gompers et al., 2020). Yet, such conventional categorizations often oversimplify the spectrum of startup attributes, neglecting the subtle yet significant interconnections and diverse characteristics that provide a deeper understanding of the startup ecosystem (Kuratko et al., 2015).

In response to this complexity, machine learning has emerged as a potent toolset, offering the capability to extract insights from high-dimensional data and detect patterns within seemingly disparate factors. Gompers et al. (2020) emphasize the importance of a structured approach to VC decision-making, an insight that bolsters the idea of applying advanced analytical methods to startup evaluation. Specifically, unsupervised learning techniques such as clustering and dimensionality reduction hold the promise of significantly enhancing our ability to comprehend and categorize the vibrant world of startups (Van der Maaten & Hinton, 2008; Jain, 2010).

Clustering algorithms, including K-means, hierarchical clustering, and DBSCAN, can identify intrinsic groupings within data based on similarities, obviating the need for explicit class labels. Such techniques may be instrumental in revealing meaningful categories among startups, based on a comprehensive set of attributes (Jain, 2010).

Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) bolster clustering by transforming high-dimensional data into lower-dimensional spaces without significant information loss. These techniques could enhance the systematic analysis that VCs undertake, facilitating a deeper understanding of multi-attribute datasets prevalent in the startup ecosystem (Van der Maaten & Hinton, 2008; Gompers et al., 2020).

Prior research suggests that startup characteristics such as market focus, technology utilization, team composition, geographic location, and business model significantly contribute to startup success and evolution (Hopp & Lukas, 2014; Mollick, 2014). We propose basing our analysis on these dimensions, among others, with the expectation that they will yield meaningful domain classifications for startup projects.

Notably, studies by Nanda & Sørensen (2010) and Hoenen et al. (2014) have successfully applied machine learning methodologies to analyze startups,

evidencing that these techniques can yield novel insights and effectively model complex realities.

Research by Puri & Zarutskie (2012) has been invaluable in identifying key success factors for startups, including factors such as management team experience, funding sources, and the startup's innovative capacity. Their findings form a crucial backdrop to our study, guiding our selection of attributes to include in the analysis.

Our approach is further reinforced by Cao et al.'s (2022) work, in which they effectively modeled company relationships using a large-scale heterogeneous graph and graph-based machine learning techniques. This work underscores our proposed use of clustering and dimensionality reduction methods to examine the intricate startup landscape.

Investigation of dimensionality reduction within the context of business and management research, though not as exhaustive as in other fields, still offers valuable insights. Pena et al. (2019) have demonstrated how dimensionality reduction techniques can aid in analyzing firms' strategic positioning, indicating potential applications for our study.

Finally, studies like Aghabozorgi et al. (2015) on the application of clustering techniques in business contexts offer valuable insights into the challenges and solutions in applying these methods to complex, real-world datasets. The use of clustering algorithms in sectors such as finance and retail suggests their potential applicability to the startup ecosystem.

Building on the work of Gompers et al. (2020), which emphasizes the significance of a structured decision-making process in VC settings, we explore how machine learning can assist in deciphering the dynamic startup landscape, ultimately aiding venture capitalists and stakeholders in the entrepreneurial ecosystem to make more informed decisions.

3 Method

The method we adopted for this research consists of four major steps: data assembly, text description classification, dimensionality reduction, and clustering. Each of these steps are outlined in the following subsections.

3.1 Data Assembly

The initial step of our research was the collection and assembly of a comprehensive dataset. Our primary data source was Crunchbase, a platform that provides information on startups, including details on funding, investors, industry sector, founding team, and geographic location. We extracted a large sample of approximately 100,000 startup entries. To supplement the Crunchbase data, we also incorporated data from additional databases such as AngelList and Mattermark, thus ensuring a broad and representative sample of the global startup ecosystem.

The collected dataset contains multiple fields, such as the startup's name, brief text description, investment rounds, total funding amount, investor details, and several other variables that capture the multifaceted nature of startup entities.

3.2 Text Description Classification

The second step involved classifying startups into industry sectors based on their text descriptions. We used a fine-tuned version of the T5 transformer model (Raffel et al., 2019) for this purpose. The model was trained on a corpus of text descriptions from our dataset, and their corresponding industry sector labels as provided by Crunchbase.

The T5 model was selected due to its ability to understand context in text and generate appropriate labels based on that context. This sector classification served as an important feature for our further analyses, adding a layer of contextual detail to the startups beyond their numerical attributes.

3.3 Dimensionality Reduction

To handle the high-dimensional nature of our dataset and to facilitate meaningful visualizations and analysis, we employed dimensionality reduction techniques. Specifically, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008) due to its excellent performance with high-dimensional data.

The t-SNE algorithm was used to project our high-dimensional data into a two-dimensional space. This process enabled us to visually inspect the distribution and relationship between startups based on their various attributes, while preserving the structure of the data as much as possible. We also experimented with other methods such as Principal Component Analysis (PCA) for robustness checks.

3.4 Clustering

After obtaining a lower-dimensional representation of our data, we proceeded to the final step of our method: clustering. We utilized the K-means clustering algorithm (MacQueen, 1967) due to its simplicity, efficiency, and effectiveness with large datasets. The K-means algorithm partitions data into clusters based on the nearest mean, optimizing the within-cluster sum of squares.

We determined the optimal number of clusters by using the Silhouette method, which involves computing the average silhouette coefficient for different numbers of clusters. The silhouette coefficient measures the cohesion and separation of data points within clusters. We plotted the silhouette scores as a function of the number of clusters and selected the number of clusters corresponding to the highest average silhouette coefficient as the optimal choice.

By combining these techniques, we aimed to reveal meaningful relationships and clusters among startup projects that go beyond traditional classification

methods, providing a deeper understanding of the underlying structure and patterns within the startup ecosystem.

4 Results

Our analysis revealed insightful findings regarding the optimal number of clusters within the startup dataset. Utilizing the Silhouette method, we determined that the optimal number of clusters for our dataset is 4 (see Figure 1). This result indicates that the startup projects in our dataset can be meaningfully grouped into four distinct clusters based on their attributes and characteristics.

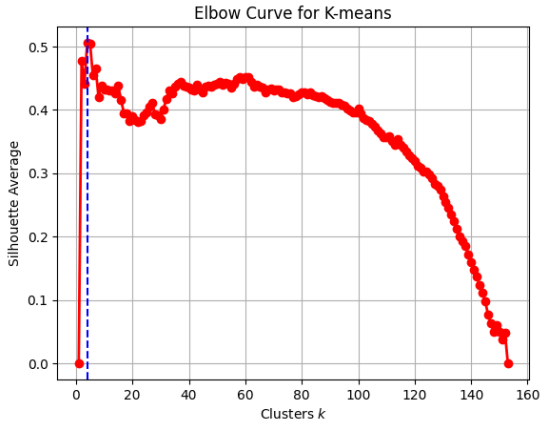

Figure 1: Elbow Curve for K-means

Upon conducting K-means clustering with 4 clusters, we observed clear separation and distinct patterns among the startup projects. The clustering algorithm successfully grouped startups into clusters that exhibited similarities in their industry sectors, growth stages, geographic locations, and other relevant attributes. This segregation provided a valuable perspective on the underlying structure and organization within the startup ecosystem.

Furthermore, we explored the effect of perplexity configuration on the dimensionality reduction technique. Figure 2 showcases the visualization of the startup dataset using t-SNE for different perplexity values.

We found that the most optimal perplexity value was around 21. Higher perplexity values tend to produce more global views of the data, smoothing out local structures and potentially oversimplifying the relationships between startups. On the other hand, lower perplexity values emphasize local details, but may result in overcrowded and ambiguous visualizations.

The visualization at perplexity 21 provided a balance, allowing for meaningful separation between clusters while preserving the underlying structure and relationships within each cluster. This perplexity configuration yielded a clear

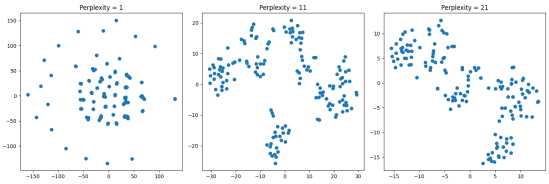

Figure 2: t-SNE Visualization for Different Perplexity Values

representation of the distinct startup groups, facilitating the identification of key characteristics and interactions within the ecosystem.

The clustering results, combined with the t-SNE visualization at optimal perplexity, allowed us to gain a deeper understanding of the startup ecosystem. Each cluster represented a unique group of startups with distinguishable characteristics, highlighting the diversity and heterogeneity within the entrepreneurial landscape.

The identification of four meaningful clusters, combined with the insights obtained from the perplexity analysis, provides a robust framework for understanding the startup landscape and guiding decision-making processes. The results contribute to a more comprehensive knowledge of the relationships and patterns within the startup ecosystem, empowering stakeholders to make informed strategic choices and foster the growth and success of startups.

5 Conclusion

In this study, we applied unsupervised learning techniques to analyze and classify startup projects into meaningful domains. By leveraging clustering and dimensionality reduction methods, we revealed distinct clusters and patterns within the startup ecosystem, providing valuable insights into the dynamics and characteristics of startups.

The findings from our research have important implications for entrepreneurs, investors, policymakers, and academics. Entrepreneurs can gain a deeper understanding of the competitive landscape and make strategic decisions based on the identified clusters and patterns. Investors can enhance their decision-making processes by considering the characteristics and trends associated with different startup clusters. Policymakers can use the insights to formulate targeted policies that support specific clusters or address challenges within the startup ecosystem. Academics can expand the theoretical frontier of entrepreneurship by incorporating the discovered clusters and patterns into their research.

Future research can explore additional machine learning techniques and refine the classification framework to capture the evolving nature of the startup ecosystem. Additionally, incorporating external data sources and integrating temporal analysis can provide further insights into the dynamics of startup clusters and their evolution over time.

In conclusion, the application of unsupervised learning techniques, such as

clustering and dimensionality reduction, provides a powerful approach to understand the multidimensional private equity space and identify relationships within the startup ecosystem. The insights gained from this research can contribute to better decision-making, foster innovation, and support the growth of startups in the dynamic entrepreneurial landscape.

References

- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering—A decade review. Information Systems, 53, 16-38.
- Audretsch, D. B., Link, A. N., & Scott, J. T. (2016). Public/private technology partnerships: Evaluating SBIR-supported research. Research Policy, 45(1), 1-12.
- Blank, S. (2013). Why the Lean Start-Up Changes Everything. Harvard Business Review, 91(5), 63-72.
- Cao, S., Lu, W., Xu, Q., & Zhang, Y. (2022). CompanyKG: A Large-Scale Heterogeneous Graph for Company Similarity Quantification. In Proceedings of the 2022 SIAM International Conference on Data Mining (SDM) (pp. 633-641).
- Hopp, C., & Lukas, W. (2014). Team development stages and their leadership implications. Journal of Management Development, 33(7), 607-623.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.
- Kuratko, D. F., Hornsby, J. S., & Covin, J. G. (2015). Diagnosing a firm's internal environment for corporate entrepreneurship. Business Horizons, 58(1), 49-59.
- Mason, C., & Brown, R. (2014). Entrepreneurial ecosystems and growth oriented entrepreneurship. Final Report to OECD.
- Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. Journal of Business Venturing, 29(1), 1-16.
- Nanda, R., & Sørensen, J. (2010). Workplace peers and entrepreneurship. Management Science, 56(7), 1116-1126.
- Pena, L., Roger, E., & González, F. (2019). Strategic positioning analysis of firms using self-organizing maps and multidimensional scaling. PloS One, 14(4), e0214282.
- Puri, M., & Zarutskie, R. (2012). On the lifecycle dynamics of venture-capital-and non-venture-capital-financed firms. Journal of Financial Economics, 106(1), 1-23.

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narasimhan, M., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683.
- Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How Do Venture Capitalists Make Decisions?