UNVEILING CROSS-SECTORAL DYNAMICS AND TRENDS: ANALYZING THE INDIAN STARTUP ECOSYSTEM THROUGH STRUCTURAL TOPIC MODELLING

ABSTRACT

India is witnessing an unprecedented growth within the startup landscape, as it continues to solidify its position as a major global centre for innovation and entrepreneurship. This research paper aims to uncover the latent sectoral topics and analyse trends in the Indian startup ecosystem using Structural Topic Modelling (STM).

Candidate ID: 28017

Abstract

India is witnessing an unprecedented growth within the startup landscape, as it continues to solidify its position as a major global centre for innovation and entrepreneurship. This research paper aims to uncover the latent sectoral topics and analyse trends in the Indian startup ecosystem using Structural Topic Modelling (STM). The analysis leverages company descriptions from LinkedIn's About section, along with a rich meta-data, of 661 startup companies, founded between 2018-2023. The study uncovers a growing multidisciplinary nature of startups among top-funded sectors, driven by digital and AI technologies, and identifies emerging trends such as the expanding presence of Ecommerce, Edtech and Agritech beyond Tier 1 cities. The findings of this study can potentially inform policymakers, investors, and entrepreneurs seeking to navigate the evolving landscape of the Indian startup ecosystem.

Introduction

According to the Economic Survey 2021-22, India has emerged as the world's third largest startup ecosystem, with over 1.25 lakh startups and 110 unicorns, attracting attention from investors across the globe for its remarkable growth and diverse range of ventures. The rapid increase in Internet penetration amongst the Indian population, coupled with targeted govt. initiatives and schemes, such as the Startup India initiative, have a played a pivotal role in boosting and unlocking the growth potential of the Indian startup industry, over the past decade. This catalysation has triggered a sudden wave of entrepreneurial activity and innovation in the country, reflected in the emergence of a diverse range of startup ventures across multiple sectors, including Agritech, Fintech, Edtech, E-commerce, etc, most of which are technologically or digitally enabled. For example, a company like DGV, which is an Integrated Dairy Fintech, Insurtech and Marketplace Platform, could actually feature across multiple sectors, such as Fintech, Agritech and Enterprise Tech. However, existing single-level classifications often overlook the nuanced and cross-sectoral themes of startups' offerings. To address this gap, this paper aims to identify latent topics across 644 startups founded between 2018-2023 and explore the occurrence of topic pairs, illustrating their multidisciplinary nature. Additionally, the study observes a positive trend in sectors with lower shares in average funding distribution, and a shift in their concentration to Tier-2 and Tier-3 cities. This analysis of sectoral themes and trends holds significance for investors, policymakers, and entrepreneurs navigating the dynamic Indian startup landscape.

Literature Review

Topic modelling and its applications

Topic modelling (TM) is certainly witnessing an unprecedented rise in interest and up-take by the social science research community, at the back of an explosive increase in text information being accessible and shared online, in this digital age. TM assumes each word follows a two-step process: iteratively assigns topic distributions to descriptions and word distributions to topics, refining until the best model is found. Unlike simple keyword counting, TM considers word context, capturing nuanced semantic relationships (Savin et. al., 2022).

Topic modelling has found a wide application across various disciplines of social sciences, particularly to understand the topical trends in political discourses, news articles, reports, and posts on social networks. For instance, Jang, et. al. (2021) used STM to analyse salient topics across news articles and magazine publications with the keyword 'startup', to explain the differences in startup discourse in US and China, based on Hofstede's cultural dimensions framework.

Startup research and topic modelling

Over the years, startups have gained significant traction in the social science research community due to their growing potential to impact societies, by driving innovation, creating jobs, and stimulating economic growth. And therefore, most of the previous research on startups have attempted to predict the success of startups or explore the factors that influence their success. For instance, Gastaud et. al. (2019) explored the role of competition and investor's network centrality in explaining the fundraising success of startups across varying stages of their growth.

Lately, a few of the contemporary studies have tapped into areas similar to our research, wherein they apply topic modelling techniques to company descriptions and draw insights relevant to the startup ecosystem. One such study by Savin et al. (2022) utilized a database of 250,000 global startup companies from Crunchbase, employing STM to reclassify company sectors into 38 topic clusters based on company descriptions. They analyzed global trends in topic prevalence across covariates like geographic location and year of establishment. Similarly, Chae et al. (2021) explored venture activities related to digital technologies using STM on company descriptions from a Crunchbase database of 133,334 US startups. They found digital technologies being recombined for innovative solutions.

Motivation/Research Gap

This paper uniquely explores trends in the Indian startup ecosystem using topic modeling, contrasting with studies on global or US-based startups. The Indian startup landscape, concentrated in Tier 1 cities, faces challenges like fragmented consumer markets and infrastructure gaps (David et al., 2021). This focus is crucial given the specific challenges shaping Indian markets, policies, and the economy. Additionally, the study stands out for using LinkedIn's About section, providing insights into how startups brand themselves and prioritize themes

Data collection, cleaning and summary

For this study, we developed a web-scraping function using R Selenium to gather data on 700 startups from Inc42 Datalabs [1], an Indian startup database. After cleaning the dataset and filtering out entries with missing or invalid LinkedIn URLs, we were left with 670 startups. The startup dataset includes a comprehensive list of features, including information on sectors, funding, founding year, headquarters, LinkedIn URLs, etc. We then used the LinkedIn URLs to extract 'About' description text for each startup, resulting in successful extraction for 616 startups. Manual retrieval was attempted for the remaining 54, with 9 having no description. Further text cleaning involved reformatting column classes, and excluding descriptions with fewer than 7 words, considering they would be uninformative. Fig 1. Illustrates the detailed flow chart for the complete data extraction process.

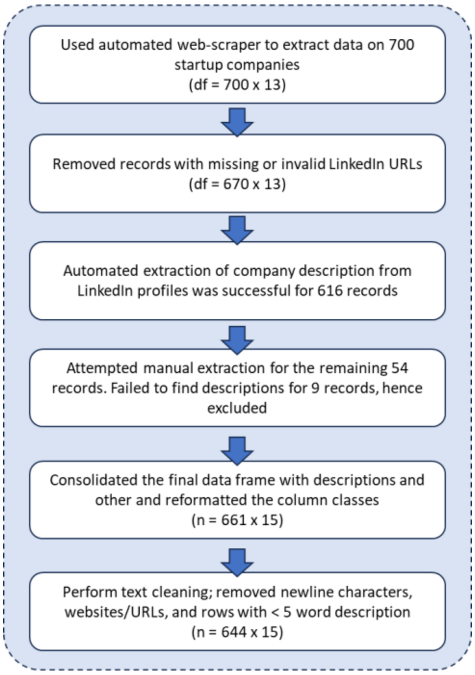
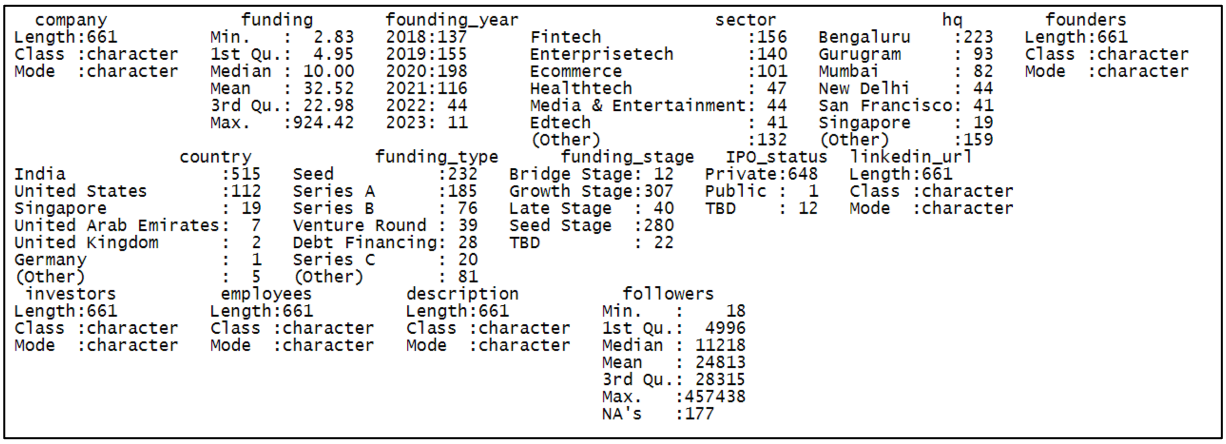

Used automated web-scraper to extract data on 700 startup companies (df = 700 x 13)
Removed records with missing or invalid LinkedIn URLs (df = 670 x 13)
Automated extraction of company description from LinkedIn profiles was successful for 616 records
Attempted manual extraction for the remaining 54 records. Failed to find descriptions for 9 records, hence excluded
Consolidated the final data frame with descriptions and other and reformatted the column classes (n = 661 x 15)
Perform text cleaning; removed newline characters, websites/URLs, and rows with < 5 word description (n = 644 x 15)

Fig. 1: Flow Chart for Data Extraction Process

The final dataset consists of 644 rows, representing data on each startup company across the 15 feature columns, including company description, funding raised, founding year, headquarter, country, funding stage, funding type, IPO status, number of investors, number of employees, follower counts on LinkedIn, etc. Fig 2. shows the R output for the statistical summary of the dataset.



```
  company             funding       founding_year                  sector                hq          founders
Length:661         Min.   :  2.83   2018:137      Fintech              :156   Bengaluru    :223   Length:661
Class :character   1st Qu.:  4.95   2019:155      Enterprisetech       :140   Gurugram     : 93   Class :character
Mode  :character   Median : 10.00   2020:198      Ecommerce            :101   Mumbai       : 82   Mode  :character
                   Mean   : 32.52   2021:116      Healthtech           : 47   New Delhi    : 44
                   3rd Qu.: 22.98   2022: 44      Media & Entertainment: 44   San Francisco: 41
                   Max.   :924.42   2023: 11      Edtech               : 41   Singapore    : 19
                                                  (Other)              :132   (Other)      :159
                country               funding_type      funding_stage    IPO_status  linkedin_url
India               :515   Seed           :232   Bridge Stage: 12   Private:648   Length:661
United States       :112   Series A       :185   Growth Stage:307   Public :  1   Class :character
Singapore           : 19   Series B       : 76   Late Stage  : 40   TBD    : 12   Mode  :character
United Arab Emirates:  7   Venture Round  : 39   Seed Stage  :280
United Kingdom      :  2   Debt Financing: 28   TBD         : 22
Germany             :  1   Series C       : 20
(Other)             :  5   (Other)        : 81
 investors           employees          description          followers
Length:661         Length:661         Length:661         Min.   :    18
Class :character   Class :character   Class :character   1st Qu.:  4996
Mode  :character   Mode  :character   Mode  :character   Median : 11218
                                                         Mean   : 24813
                                                         3rd Qu.: 28315
                                                         Max.   :457438
                                                         NA's   :177
```

Fig. 2: Statistical summary of the dataset (R output)

On further visualizing the distribution of our startup companies across sectors, we observe that Fintech emerges as the most prevalent sector in the dataset, with a 24% share amongst startups in consideration, followed by Enterprisetech (21%) and Ecommerce (15%), which remains consistent with the general trend in the Indian startup landscape, where these three sectors have been repeatedly recognized as the top-funded and most competitive startup sectors in the recent past[6]. Primarily attributed to the combined effect of the recent push to digital India, wide adoption of digital payments enabled by UPI (Unified Payments Interface) and the rising trend in the per capita GDP.

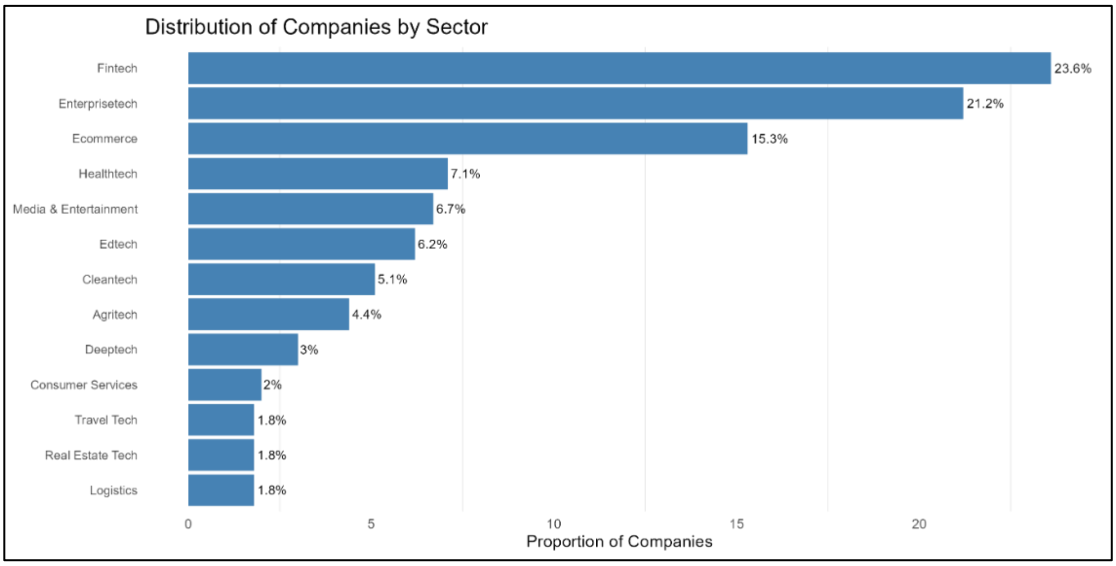


Fig. 3: Distribution of companies by sector

Next, we visualize the word-length distribution of the company descriptions and observe a right skewed distribution with half of the companies having words ranging between 25-105 in their descriptions, while the other half exceeds the mean of 103 words, with a fraction of them even going beyond 200 & 300 words. This bodes well for our choice of topic modelling method, considering STM assumes there to be multiple topics in a document. And therefore, very short texts would have been rather uninformative.

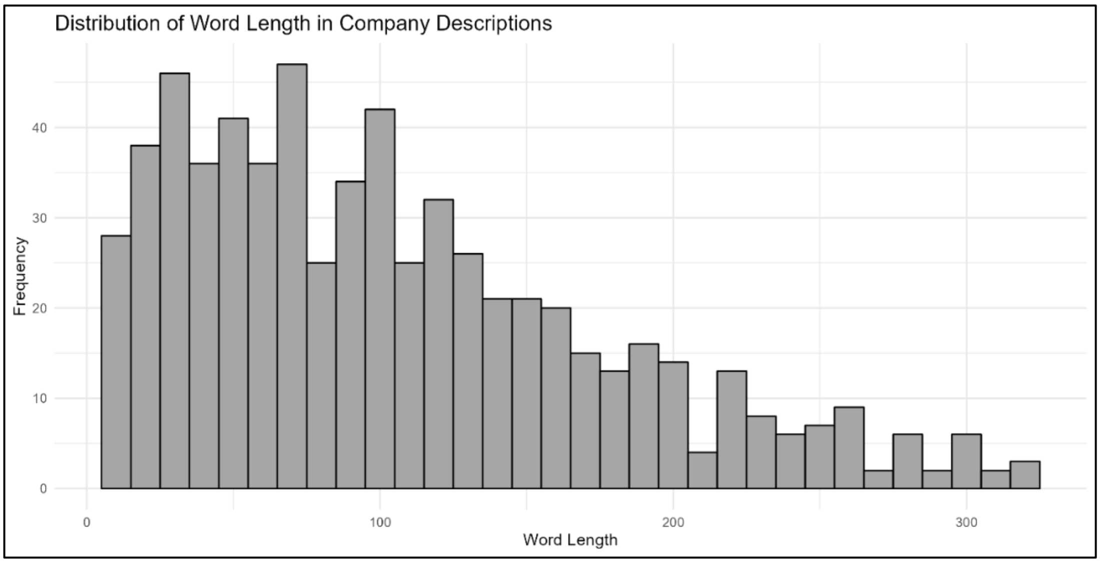


Fig. 4: Distribution of word length in company descriptions

Methodology

To identify the latent topics across the company descriptions, we employ the Structural Topic Model (STM) package in R, by Roberts et al. (2019), which is a generative model that allows for the identification of topics in a corpus of text data, while also accounting for the document-level covariates that may influence the distribution of topics across documents or the distribution of words across topics (Roberts et al., 2014). Therefore, STM is particularly appropriate for our context as it not only allows us to identify the latent topics but also enable us to observe any trends in topic prevalence based on geographical locations, year of establishment etc.

For text preprocessing, we utilize the 'quanteda' package to convert the descriptions into a corpus, applying standard steps such as tokenization, lowercasing, and removal of stopwords, punctuations, URLs, and symbols. A document frequency matrix (dfm) is created with a minimum term frequency of 10, then converted into an STM-compatible format. To determine the optimal number of topics (k), we consider model performance based on three criteria: Residuals (measure of how good the topics fit the data), Semantic coherence (degree of co-occurrence of most probable words in a topic), and Exclusivity (weighted harmonic mean of the word's rank in terms of exclusivity and frequency). Based on the diagnostic outputs, we select the optimal number of topics to be k=17, that best balances the trade-off between the three criteria.

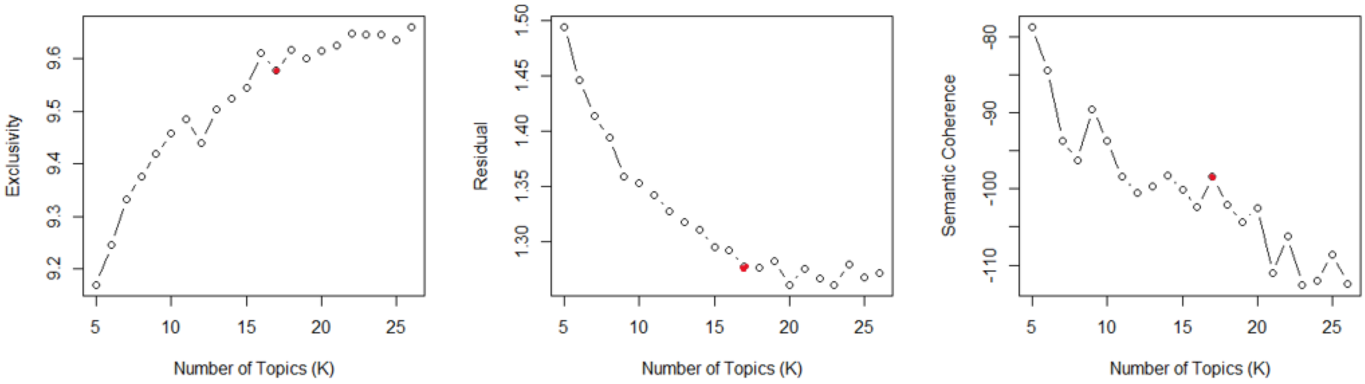
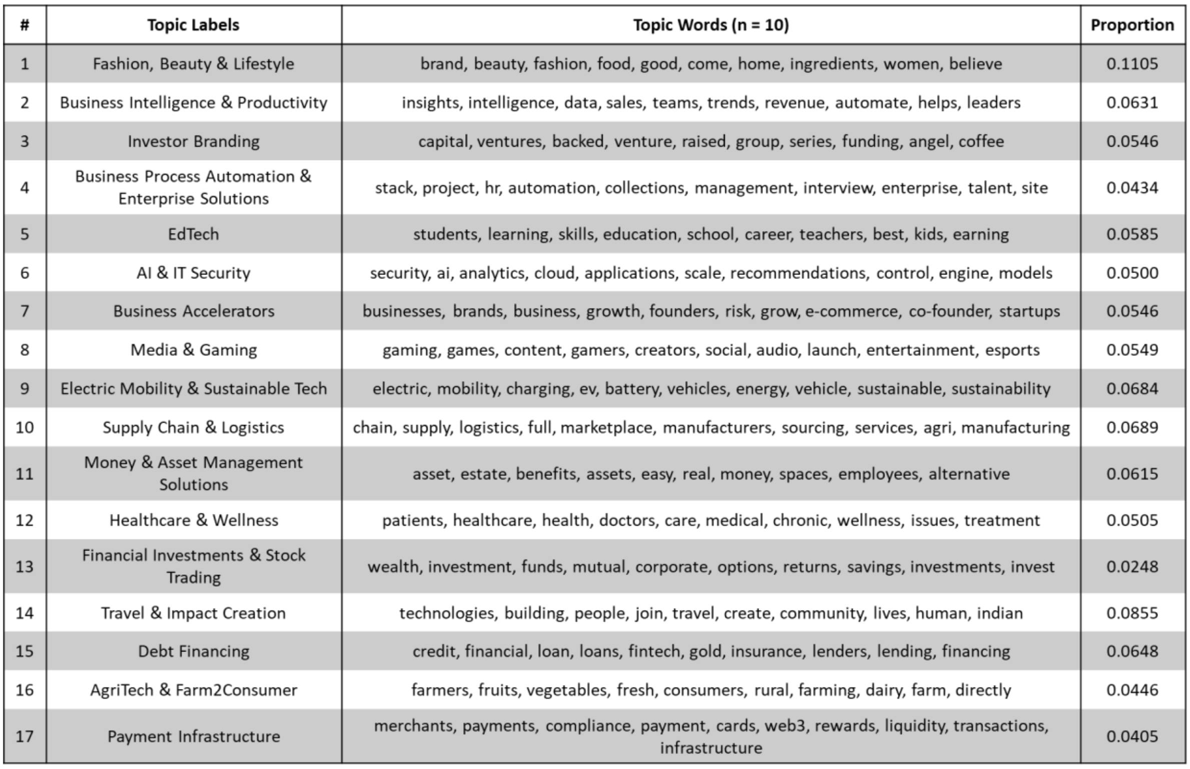


Fig. 5: Diagnostic output for Exclusivity, Residual, and Semantic Coherence.

Results and discussions

Topic identification and prevalence

The applied STM model generates meaningful topics and is able to classify every company description into one of the 17 topics, without leaving out any isolates. Table 1, offers details on the 17 topics identified based on the company descriptions obtained from the About section of their LinkedIn profiles.



#	Topic Labels	Topic Words (n = 10)	Proportion
1	Fashion, Beauty & Lifestyle	brand, beauty, fashion, food, good, come, home, ingredients, women, believe	0.1105
2	Business Intelligence & Productivity	insights, intelligence, data, sales, teams, trends, revenue, automate, helps, leaders	0.0631
3	Investor Branding	capital, ventures, backed, venture, raised, group, series, funding, angel, coffee	0.0546
4	Business Process Automation & Enterprise Solutions	stack, project, hr, automation, collections, management, interview, enterprise, talent, site	0.0434
5	EdTech	students, learning, skills, education, school, career, teachers, best, kids, earning	0.0585
6	AI & IT Security	security, ai, analytics, cloud, applications, scale, recommendations, control, engine, models	0.0500
7	Business Accelerators	businesses, brands, business, growth, founders, risk, grow, e-commerce, co-founder, startups	0.0546
8	Media & Gaming	gaming, games, content, gamers, creators, social, audio, launch, entertainment, esports	0.0549
9	Electric Mobility & Sustainable Tech	electric, mobility, charging, ev, battery, vehicles, energy, vehicle, sustainable, sustainability	0.0684
10	Supply Chain & Logistics	chain, supply, logistics, full, marketplace, manufacturers, sourcing, services, agri, manufacturing	0.0689
11	Money & Asset Management Solutions	asset, estate, benefits, assets, easy, real, money, spaces, employees, alternative	0.0615
12	Healthcare & Wellness	patients, healthcare, health, doctors, care, medical, chronic, wellness, issues, treatment	0.0505
13	Financial Investments & Stock Trading	wealth, investment, funds, mutual, corporate, options, returns, savings, investments, invest	0.0248
14	Travel & Impact Creation	technologies, building, people, join, travel, create, community, lives, human, indian	0.0855
15	Debt Financing	credit, financial, loan, loans, fintech, gold, insurance, lenders, lending, financing	0.0648
16	AgriTech & Farm2Consumer	farmers, fruits, vegetables, fresh, consumers, rural, farming, dairy, farm, directly	0.0446
17	Payment Infrastructure	merchants, payments, compliance, payment, cards, web3, rewards, liquidity, transactions, infrastructure	0.0405

Table 1: Topics generated from STM model

The Topic Words column displays the top 10 words, identified using the FREX method, favoring frequent and exclusive words within a topic. Topic labels were manually assigned based on FREX words and a review of top 10 descriptions. Proportions indicate average topic prevalence across all descriptions. Notably, topics like Impact Creation and Investor Branding may not reflect business activities but rather textual features or branding choices, emphasizing community impact or drop names of prominent investors, to gain more traction from signalling effect. Fig. 6 highlights Fashion, Beauty & Wellness as the most prevalent topic, likely due to e-commerce operations. Followed by Travel & Impact Creation follows, possibly due to the non-sector specific identity of the topic.

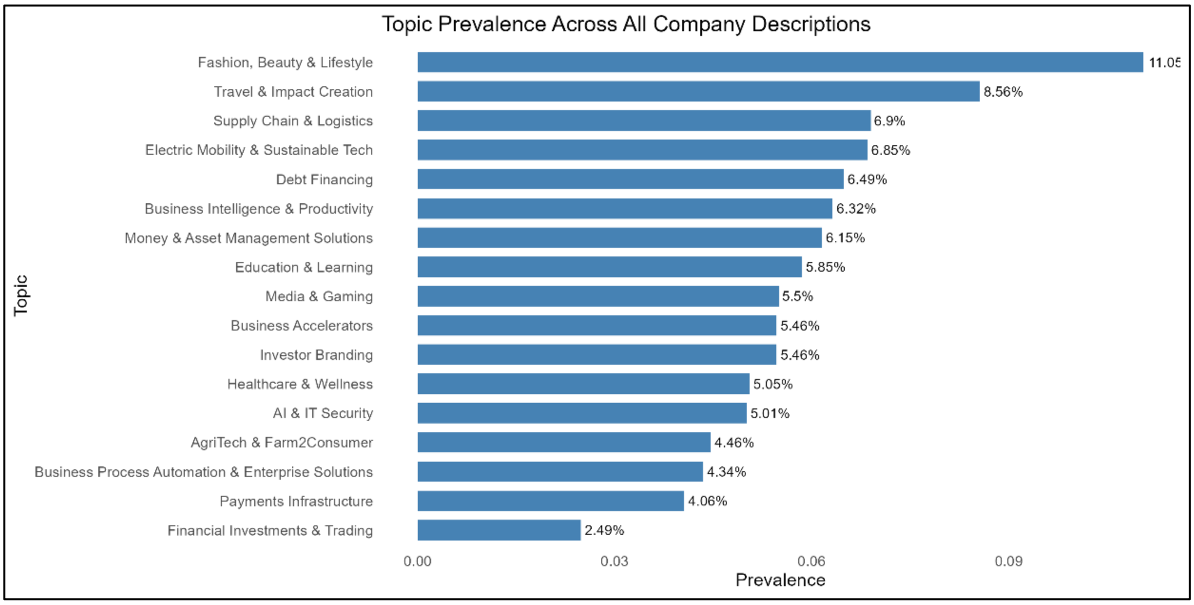


Fig. 6: Topic prevalence across company descriptions

Startup multidisciplinarity in the STM classification

STM generates non-exclusive topic classifications, allowing descriptions to encompass fragments of multiple topics, reflecting the cross-sectoral or multidimensional nature of company offerings or branding strategies. Analysing the average number of topics per description reveals this multidisciplinarity. We plot the frequency distribution of topics per startup, considering topics with a prevalence threshold of 5% to be substantial, avoiding misinterpretation of insignificant word features as topic signals. The results are depicted in Fig. 7, and it indicates that majority of the descriptions include topics in the range of 2-5 topics, with an average of 3 topics (mean = 3.5).

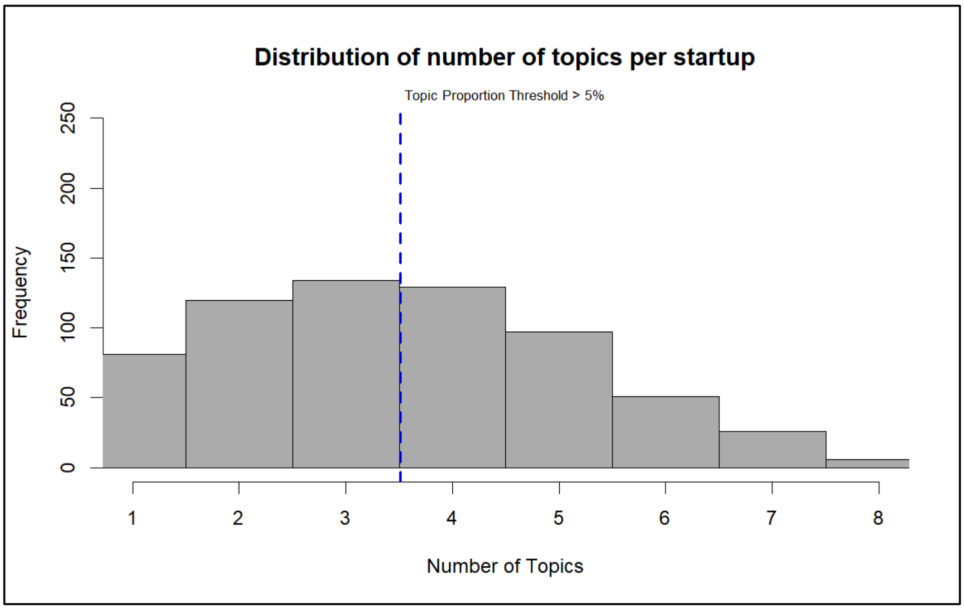


Fig. 7: Distribution of number of topics per startup

Further, on plotting (Fig. 8) the average number of topics per startup description across the years of establishment, we observe a marginal increase in the number of topic occurrences in startup descriptions, over the last 5 years. This seems to be an intuitive evolution owing to the rise in tech enabled or digitally powered startup ventures in India, which by default cuts-across a minimum of two sectors; the technology sector and the functional domain of the business, thus pushing up the average.

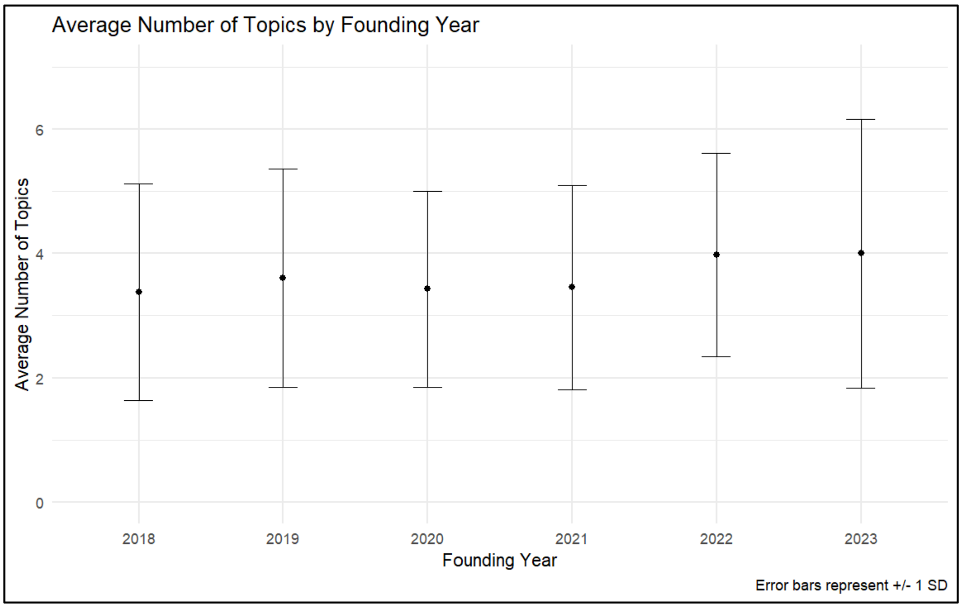


Fig. 8: Average number of topics per startup by founding year

Next, we visualize the percentage overlap between generated topics and sector classifications, to see how many latent topics a particular sector spans across, revealing the unexplored opportunity to understand the dynamics of a startup sector across nuanced sub-sectors and themes. Notably, Fintech, Enterprisetech, and Ecommerce, India's highly funded sectors, exhibit significant sectoral depth, reflected in their strong overlaps across multiple STM topics (Fig. 9). This can possibly be attributed to the saturated and evolved nature of these sectors, owing to the relative ease of fund availability and wide spread diffusion of ICT technologies in these sectors. For the purpose of observing the overlaps, we excluded Topic 3 and Topic 14, as their established non-sector specific identity would have caused bias in the interpretation.

Interestingly, unlike the expected overlap of Enterprisetech startups across Topics 6, 4 and 2, which are all targeted towards automating business functions and enhancing productivity through AI and digital technologies, the sector also shares an overlap with Topic 10: Supply Chain and Logistics, highlighting the cross-sectoral nature of buiness activities. For instance, ‘Assiduus Global’ offers full-stack tech-enabled middleware for digital distribution, inventory planning and supply chain management for brands, and therefore finds itself at the intersection of both supply chain and enterprisetech sectors.

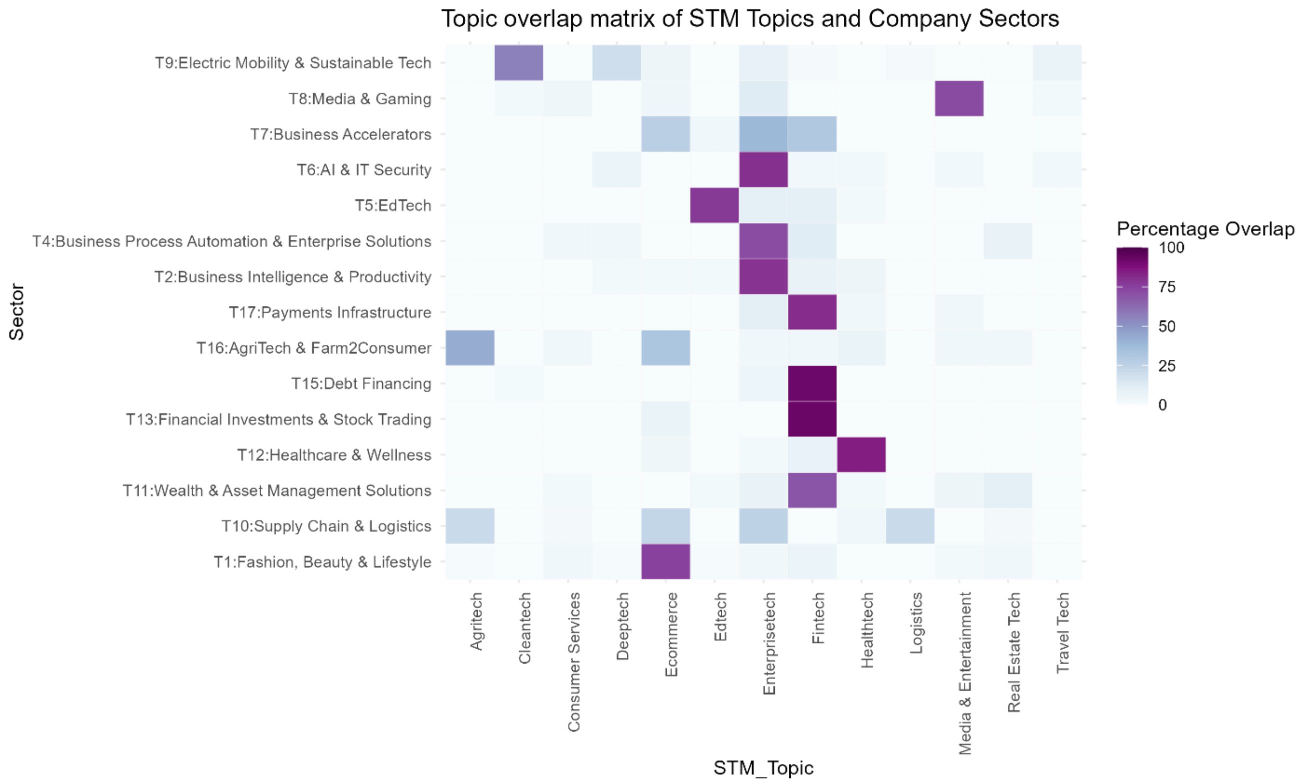


Fig. 9: Topic overlap matrix of STM topics and company sectors

Finally, Fig. 10 depicts the correlation between generated topics, with colors indicating the strength of correlation. Blue points suggest a likelihood of topic co-occurrence due to positive correlation. Apart from intuitively complementary clusters, 'EdTech' and 'Media & Gaming' exhibit a strong tendency to co-occur, reflecting India's emphasis on educational transformation through gamification and interactive media (Kumar et al., 2021). These sectors primarily target children and young adults, prompting exploration of common applications like gamified learning and interactive e-learning content. Startups such as Lido and Creative Galileo offer gamified learning solutions.

Additionally, the positive correlation observed between 'AgriTech & Delivery Services' and 'Supply Chain & Logistics' underscores their complementary relationship. AgriTech startups in India leverage IT and digital technologies to streamline supply chains and integrate fragmented marketplaces, addressing challenges in the agriculture sector. For instance, 'Fasal' and 'Wheelocity' focus on optimizing supply chains for better market prices and minimizing food wastage, respectively.

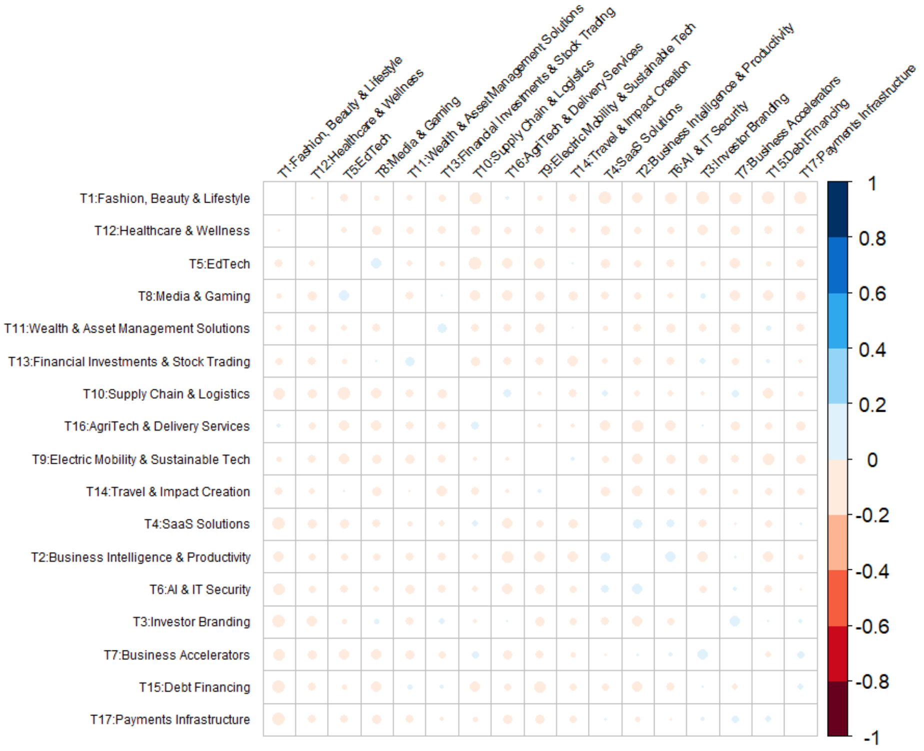


Fig. 10: Co-occurrence matrix between STM Topics

Temporal and geographic trends in topic prevalence

Finally, we analyse the effect of covariates on the variation of topic prevalence among startup descriptions. We provide results from regression analysis of the proportion of topics, dependent on our covariates of interest, such as the founding year of the startups and tiered classification of the headquarter city in India. We effectively fit a linear regression model for all the 17 topics (indexed by k), as follows:

$$\textit{Topic Prevalence}_k \ \sim \ \textit{Constant}_{k\,+}\,\textit{Founding Year} + \textit{City Tier} + \textit{Residual}_k$$

For understanding the effects of covariates, it would be justified to place our focus on topics which have a low average share in the funding distribution within our sample, with the aim to better understand the trends across sectors with a higher future growth potential. From Fig. 11, we identify the topics with averages at the lower spectrum of the fund share, such as Topic 1, 2, 4, 5, 9, 12 and 16.

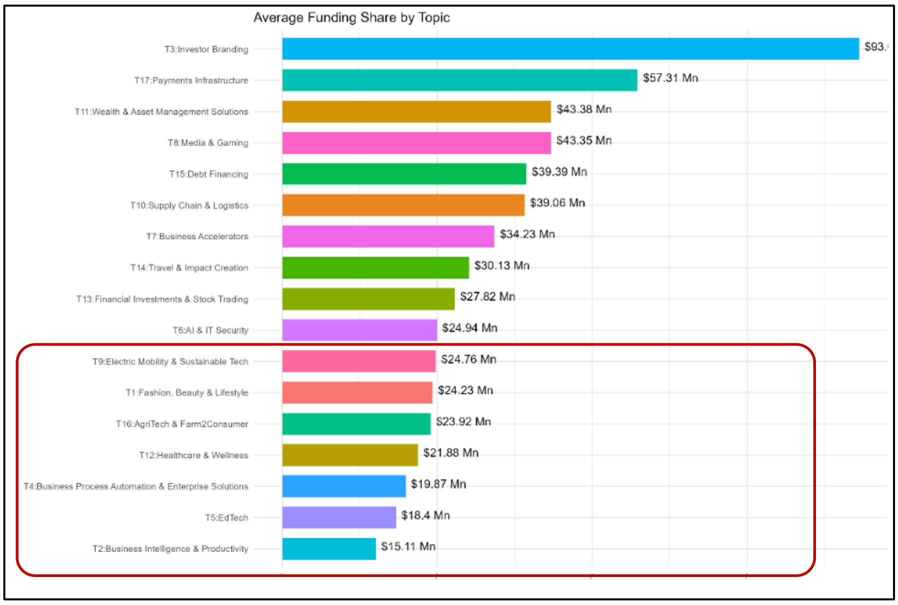


Fig. 11: Average funding share by topic

Despite a 62% decline in startup funding in India during FY23, linear trends in topic prevalence with respect to founding year show upward trajectories for most sectors, indicating growth potential. EV & Sustainable Tech and Enterprise Tech exhibit promising growth due to increased adoption of generative AI and SaaS technologies, further accelerated by the pandemic. However, Agritech and Ecommerce sectors display a negative trend, consistent with a Fortune India report indicating a 45% decline in investments in agri-tech startups between FY22 and FY23, due to hikes in global interest rates and the heightened investor caution. It also suggests that the sector is perhaps maturing, with the bulk of the venture in this space now in growth (Series A&B) or late stages (Series C and beyond).

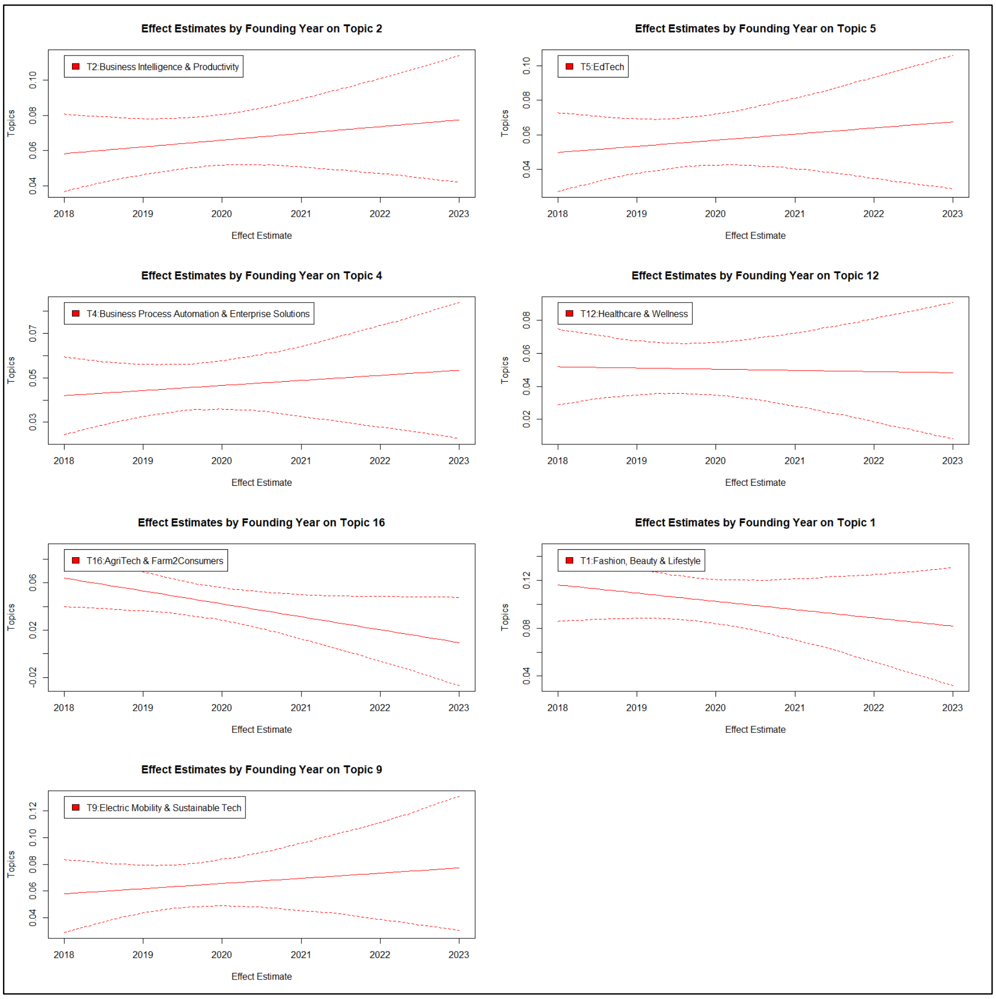


Fig. 12: Effect estimates on topic prevalence by founding year

Lastly, we attempt to understand the prevalence of topics across the different tiers of cities, specifically for startups headquartered in India. Startups in India are primarily concentrated in Tier 1 cities, particularly in IT-enabled sectors like ecommerce, fintech, and enterprisetech. However, untapped potential lies in Tier 2 and Tier 3 cities due to increasing digital inclusion, rising consumer demand, and a growing pool of entrepreneurial talent. Despite challenges such as limited ICT infrastructure and funding, these cities offer opportunities for startups to expand into emerging markets and benefit from government initiatives (Dharish, et.al., 2020).

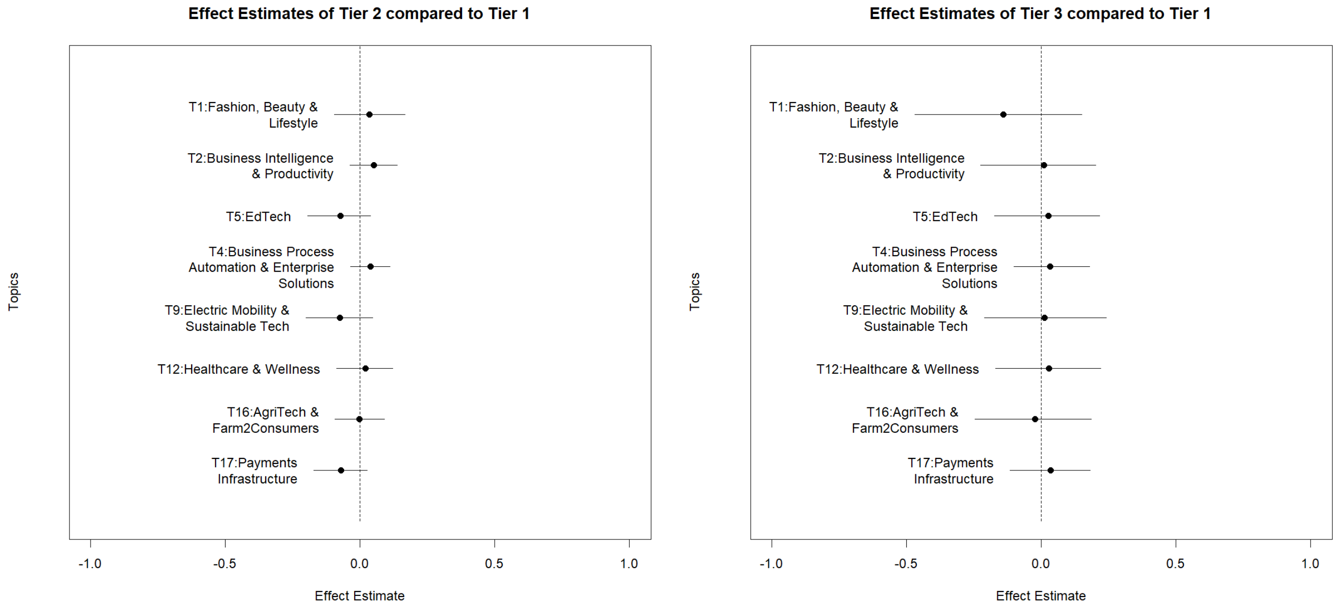


Fig. 13: Effect estimate on topic prevalence by city tiers

Notably, the prevalence of topics such as Electric Mobility & Sustainable Tech, EdTech, and Payments Infrastructure in Tier-2 cities aligns with notable startup success stories like DeHaat (AgriTech from Patna) and RazorPay (Payment Gateway Tech from Jaipur). Similarly, in Tier-3 cities, AgriTech and Ecommerce are more prevalent, reflecting accelerated ecommerce adoption facilitated by increased internet penetration and mobile internet usage. Rural India surpassed urban areas in internet users in 2019, with around 227 million active users, driving ecommerce growth in Tier 2 and Tier 3 cities, which now contribute approximately 66% of India's total online consumer demand (Kakkar, 2020). During Flipkart's Big Billion Day sale, 52% of visits came from Tier 3 and beyond cities, highlighting their significant potential in driving ecommerce expansion and sales growth.

Limitations

Despite leveraging a dataset of 644 startup companies, this study acknowledges its limited scope and suggests scaling up the analysis to include more startups and additional years for a comprehensive understanding. Due to this limitation, many findings regarding the effect estimates of topic prevalence across covariates lack statistical significance, introducing uncertainty into the observations. However, the study's consistency with broader trends in the Indian startup ecosystem enhances its credibility.

Conclusions

This study leveraged LinkedIn profiles to uncover prevalent topics and cross-sectoral themes among Indian startups. Over the past five years, startups have become more multidisciplinary due to the diffusion of digital and AI technologies, leading to the emergence of niche sub-sectors. Key sectors like Fintech, Ecommerce, and EnterpriseTech exhibit significant sectoral depth, giving rise to niche sub-categories, such as Payments Infrastructure, Digital Lending, and Wealth Management from Fintech, and SaaS Solutions, Business Intelligence, and IT security from EnterpriseTech. Additionally, complementary topics, such as Media & Gaming and Edtech or AgriTech and Supply Chain, often co-occur in company descriptions, suggesting new business avenues for entrepreneurs and investors to explore. Moreover, the study highlights temporal and geographical trends in the Indian startup ecosystem, revealing a shift beyond Tier 1 cities. Sectors like AgriTech, Ecommerce, EV & Sustainable Tech, and Edtech are gaining traction in Tier 1 and Tier 2 cities due to increased internet penetration and demand. This indicates the evolving landscape of the Indian startup ecosystem, with Tier 1 cities no longer dominating. These findings offer valuable insights into the future of entrepreneurship in India, informing strategic decisions for stakeholders in the startup ecosystem.

References

1. 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
2. David, D. (2020). The startup environment and funding activity in India. Tokyo: Asian Development Bank Institute (ADBI).
3. David, D., Gopalan, S., & Ramachandran, S. (2021). The startup environment and funding activity in India. In Investment in startups and small business financing (pp. 193-232).
4. Fortune India. (n.d.). Agri-tech startups saw more deals, less investments in FY23. Fortune India. Retrieved from https://www.fortuneindia.com/macro/agri-tech-startups-saw-more-deals-less-investments-in-fy23/.
5. Gastaud, C., Carniel, T., & Dalle, J. M. (2019). The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage. arXiv preprint arXiv:1906.03210.
6. Inc42. (n.d.). Here's how top startup sectors performed on the funding front in 2023. Inc42. Retrieved from https://inc42.com/features/heres-how-top-startup-sectors-performed-on-the-funding-front-in-2023/.
7. Jang, J., Kim, B., Lee, K. R., & Kim, J. H. (2021). A cross-cultural comparative study on the startup discourse in 2000–2019 between United States and China. In 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-5). Seoul, Korea (South). https://doi.org/10.1109/IMCOM51814.2021.9377432
8. Kakar, S. (2020). Tier 2 and Tier 3 cities: The growth engines of India. International Journal of Education, Modern Management, Applied Science & Social Science (IJEMMASSS), 2(4), 193-197.
9. Kumar, M. D. (2023). A study on education system to utilize new techniques for digital education in India. International Journal of Multidisciplinary Studies, 7(4), 78-86. https://doi.org/10.21917/ijms.2023.0267
10. Savin, I., Chukavina, K., & Pushkarev, A. (2023). Topic-based classification and identification of global trends for startup companies. Small Business Economics, 60, 659–689. https://doi.org/10.1007/s11187-022-00609-6