Kalpa Publications in Computing

Volume 19, 2024, Pages 134–143

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

Enhancing Startup Resource Discovery: A Machine Learning Approach with Vector Embeddings

Varun Kumar V[1], Ruthvik S[2] and Jinu Sophia J[3]

[1]Rajalakshmi Engineering College, Chennai, India
201401054@rajalakshmi.edu.in
[2] Rajalakshmi Engineering College, Chennai, India
201401039@rajalakshmi.edu.in
[3] Rajalakshmi Engineering College, Chennai, India
jinusophia.j@rajalakshmi.edu.in

Abstract

AI has emerged as a transformative force for startups across various industries. It offers automation, data-driven decision-making, personalization, and predictive analytics, enabling startups to improve efficiency, gain a competitive advantage, and scale their operations. The startup resource dashboard application uses Machine learning to scrape data from a variety of sources, such as government websites, industry publications, and social media. The application extends beyond simple data gathering by incorporating machine learning to craft an advanced matching algorithm. It uses Large Language Models trained exclusively on data relevant to the current economic information and other data that helps the startups to take data driven decision. Furthermore, vector embeddings are employed to enhance the context and relevance of the generated responses. Encoding the words and phrases in the High-dimensional vectors, the model gains a better undersign of the startup eco-system facilitating more accurate and insightful recommendations. By bridging the gap between Advanced AI technologies and specific needs of startups, this methodology improves the innovation and success within the startup environment.

1 Introduction

Startups often face challenges in finding the resources they need to succeed. This can include funding, talent acquisition, workspace, logistics, and mentorship. The Startup Ecosystem has witnessed an explosion of creativity and entrepreneurial endeavors. The global increase in the number of innovative ideas has given rise to an increasingly competitive landscape as written (Maradi, 2023). As a result, entrepreneurs often struggle to secure the necessary resources and support to turn their visions into thriving businesses as written (Cohen et al., 2019). This situation necessitates a tailored solution that can help startups efficiently manage their operations and access crucial resources.

Researches in Sweden conducted a study on emerging startups of the country within the year 2022 – 2023 They conducted interviews with eight startups and analyzed online data. Challenges include limited context awareness, skills, data constraints, costs, overreliance, and ethics. Opportunities include enhanced efficiency, knowledge acquisition, user experience, cost- effectiveness, and innovation. AI tools employed include Language Models, Data Generators, Automatic Speech Recognition, and more. They concluded that startups that used the power of AI and ML have increased productivity and efficiency even though they had limited manpower.

Startups represent a significant driving force behind innovation and economic growth in today's global landscape. They continually disrupt traditional industries, introduce novel products and services, and create employment opportunities. However, startups also face a variety of obstacles on their path to success, including but not limited to limited access to capital, insufficient networks, and difficulties in navigating the intricacies of business operations. This methodology is devoted to addressing these challenges through the development and implementation of a dedicated application designed to cater to the unique requirements of startups

2 Literature Survey

The application of artificial intelligence (AI) in various industries, particularly in startups, has been a subject of significant interest and research in recent years. AI's transformative potential in startups has been widely recognized, offering benefits such as automation, data-driven decision-making, personalization, and predictive analytics This has enabled startups to enhance efficiency, gain a competitive edge, and scale their operations more effectively.

The application of artificial intelligence (AI) in various industries, particularly in startups, has been a subject of significant interest and research in recent years. AI's transformative potential in startups has been widely recognized, offering benefits such as automation, data-driven decision-making, personalization, and predictive analytics This has enabled startups to enhance efficiency, gain a competitive edge, and scale their operations more effectively.

In the research titled “Successfully Organizing AI Innovation Through Collaboration with Startups”, Jana Oehmichen and Alexander Schult discuss about the different hurdles faced in implementation of AI in startups. Six different AI use cases implemented by two different companies are studied and the challenges are identified as written (Oehmichen, Jana et al., 2023).

Through the development of sophisticated algorithms and databases, startups can leverage these technologies to navigate the complex landscape of available resources more effectively, ultimately driving innovation and growth within the startup ecosystem

.

3 Research Gap

India's startup ecosystem has witnessed remarkable growth and innovation in recent years as written (Maradi, 2023). With an ever-expanding pool of young entrepreneurs, a thriving tech industry, and government initiatives like "Startup India," the country has become a hotbed for startup activity. However, there are still significant research gaps and unexplored areas in the domain of supporting Indian startups.

Access to funding mechanisms: Despite the surge in startup activities, there remains an ongoing challenge regarding access to funding, especially for early-stage startups as written Bessen et al., 2022). Existing research has touched upon this issue, but more in-depth analysis is needed to understand the evolving dynamics of funding sources as written (De Cremer et al., 2023) and the impact of government policies and initiatives like the Atal Innovation Mission.

Policy implementing and impact: While "Startup India" and other government policies have been introduced to foster entrepreneurship, there is a lack of comprehensive studies examining the actual implementation and effectiveness of these policies as written (Puapongsakorn & Brazdeikyte, 2023).. Further research can evaluate the impact of these policies on startup growth, innovation, and job creation.

Gender Disparities: Research on gender disparities within the Indian startup ecosystem remains limited. More studies are needed to explore the challenges faced by women entrepreneurs, the impact of gender biases, and strategies to promote inclusivity and diversity in the startup space.

Rural and Semi Urban Entrepreneurship: Most research has primarily focused on urban startup hubs like Bengaluru and Delhi. Research gaps exist in understanding the unique challenges and opportunities for startups in rural and semi-urban areas. Investigating the potential for rural entrepreneurship and the role of technology in bridging the urban-rural divide is crucial.

4 Proposed Methodology

The proposed methodology involves collecting data from diverse sources such as government websites and social media as written (Bessen et al., 2022). We utilize vector embeddings to convert the government data, storing it in a vector database. Leveraging language models, we efficiently fetch data, preprocessing it for quality and extracting relevant features as written (Pandiaraja et al., 2022), (Ghina

& Sinaryanti, 2021). We'll then develop a machine learning model to create an advanced matching algorithm considering factors like developmental stage and industry.

After training and evaluation, the model will integrate into the startup resource dashboard application for seamless access. Rigorous testing, user feedback incorporation, and iterative improvement will ensure system effectiveness. Finally, deployment and maintenance plans will guarantee ongoing support and responsiveness to startups' evolving needs, aiming to provide a robust resource matching solution.

5 Architecture

This architectural diagram provides an overview of the project. The modules and the overall workflow are described here. Shows the key components of suggested real time Data storing and analysis module as written (Gui et al., 2020). The diagram explains the generalized workflow of the system by considering the standard assumptions.

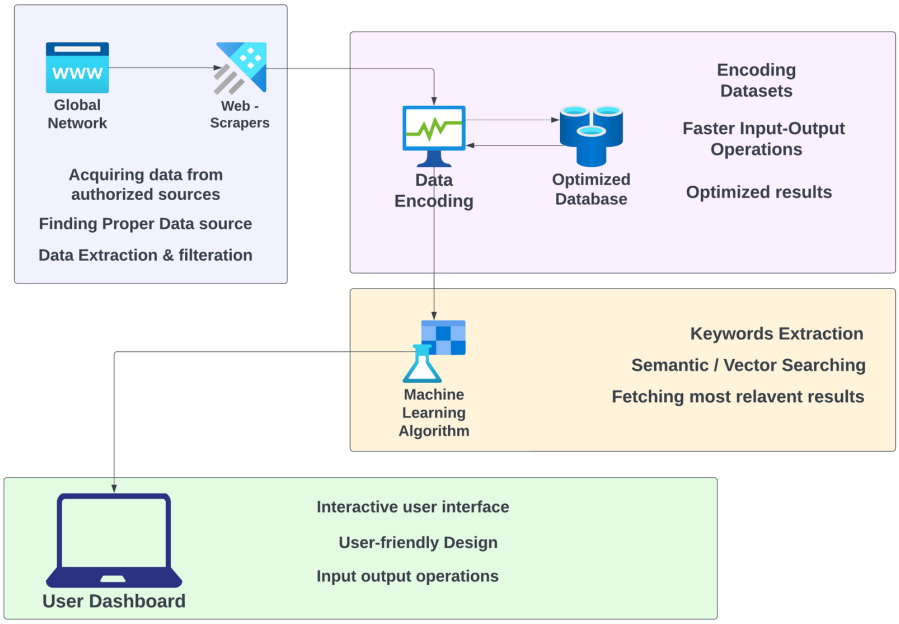

Figure 1: General Architecture diagram Explaining the major components of the application.

5.1 Data Acquisition and Preprocessing

Data Acquisition, Preprocessing and Text Extraction Further to the NLP context, lets elaborate on the initial phases of the process which is data acquisition, preprocessing and text extraction. Data acquisition is about the process of collecting or acquiring datasets from various sources such as websites, databases and files. These steps form the base foundation for the further processes. After

collecting all the necessary data, it is time to preprocess it such that the reconstruction turns the data into a structured, simple and easy to interpret format which best suites the requirements for the analysis.

This also involves filtering out the noisy data (unusual data) that exists in the datasets. Techniques such as lowercasing, stop word removal and stemming ensure that there is an uniformity within the data. Finally, text extraction focuses on retaining relevant information with methods such as named entity recognition, parts of speech tagging and sentiment analysis to uncover the meaning and tones within the text. These stages together build up for the groundwork in NLP.

Several studies explore using Machine Learning to predict the Success of Information Technology Startups (SIT). Though this approach seems promising, the main challenge lies in identifying the important factors that are responsible to determine the success of a startup. A recent study proposed a systematic method incorporating 79 success factors and various ML algorithms. This method achieved promising results, with accuracy reaching 88%. As written (Vasquez, Edilberto & Santisteban, José & Mauricio, David. (2023).)

5.2 Embeddings and Vectorization

Embedding and vectorization are the core concepts in NLP that talks about the representation of documents and words in a numerical format which is best suitable for machine learning algorithms to interpret and process. Embeddings talk about the vector representation relationships between different words and different phrases. Deep learning models like Glove or BERT are suitable for helping machines to learn about the patterns and placing words with similar meanings close to each other.

Whereas vectorization refers to the process of converting text data into numbers. Methods such as bag of words are used in this process. After the vector representations for a text data is produced, storing them efficiently such that access time and space are not compromised highly can become a challenge. A specialized vector database also known as Vector DB was designed for this very purpose. These conversions are essential when there is a requirement for similarity match checking. In conclusion, the concepts are quite excellent for converting and warehousing the texts for an increased operational performance on the texts.

5.3 Large Language Model

Large Language Models also called LLMs for short are transformers that further push the levels of natural language processing. Transformers have become the basis for many NLP models including BERT, GPT and other AI models. LLMs are often trained with a very large datasets typically the size of data lakes. These LLMs have an extra ordinary capacity for linguistic pattern understanding, text generation and logical reasoning. These models are trained on a vast variety of data. This makes the model's levels of understanding prompt texts boost up. These LLMs use self-supervised learning to learn about the new datasets that are given to it. This means that the data can be either raw, unlabeled or labelled data

One such popular LLM is Mistral 7B. Despite having a smaller size of 7.3 billion parameters, which is significantly smaller than other LLMs, it performs exceptionally good when compared with these models. Due to these aspects, Mistral is able to provide good results with limited compute power.

5.4 Prompt

Prompt is an input that is served to a LLM model or system that consists of instructions or requests that specify the need for producing a specific output that is expected by the user. The prompt directly given by users most likely will not be of an appropriate format required for the model to be able to process it and so it is structured into a set of sentences that directs the AI model or the LLM to perform the desired set of tasks the user has requested for to achieve the primary goal of producing an output. The quality of the prompt determines the accuracy and relevance of the expected output. Text translation, completion, summarization and creative writing are some of the applications of using this concept.

5.5 Storing Vector Data

After the vector representations for a text data is produced, storing them efficiently such that access time and space are not compromised highly can become a challenge. A specialized vector database also known as Vector DB was designed for this very purpose.

Unlike traditional relational databases that use tables with rows and columns, vector databases store data as vectors. Each vector has a fixed number of dimensions, which can range from tens to thousands depending on the complexity of the data.

ChromaDB is used in this approach to effective store the vectors that are created. It is an opensource vector database that is designed specifically to work with Large Language Models. It also offers easy integrations with programming languages like python and javascript. ChromaDB can operate in-memory for faster data processing and prototyping. It also allows data to be persistent using tools other dependencies

5.6 Conversation Chain Building

The Conversation Chain Building commonly referred as 'LangChain' is about building contextually linked series of dialogues in a conversation. It is about a smooth shift from one topic to another. This creates a conversational flow that cross relates different subjects from the same conversation.

LangChain is crucial in AI as a conversation may lead to questioning of the subjects or cross questioning two or more subjects together. This process involves the model to be trained with contextual understanding and the capability to trigger responses that do not deviate the conversation from the actual flow. And so the continuity of the conversation is maintained. Implementing a successful LangChain is to produce human like conversation to improve the product's / service's experience. Thus, usage of LangChain enhance the experience towards the users for a better usage.

6 Discussions

In this section, we interpret the key findings from our research on startup support in India. We discuss the significance of the results in the context of the challenges and opportunities faced by Indian startups. The findings highlight the areas where support is most needed and offer insights into the startup ecosystem. We delve into the specific challenges and opportunities faced by startups in India. We analyze the root causes of these challenges, whether they are related to regulatory issues, access to capital, market dynamics, or other factors.

Assessing the effectiveness of the learning process within the incubator presents an opportunity for conducting research using a mixed-methods approach. Various factors that impact measurements, including the capacity of each startup to onboard employees, the socioeconomic implications, and the environmental effects, can be incorporated into quantitative data collection to yield more comprehensive findings.

7 Results

The project has been implemented using Python to handle the core infrastructure. Streamlit library is used to develop basic chat interface to enable users to interact with the LLM. The application runs LLM models on local hardware rather than using models hosted on cloud. The performance may vary based on the processing power of the various devices.

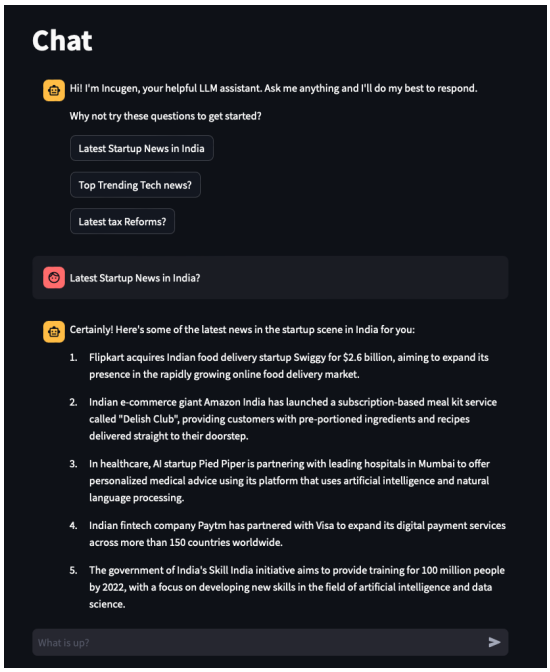

Figure 2: Chat Interface using Streamlit that serves as an interface between user and the AI-Model

An Interactive Chat Module to enable the users to interact directly with the Ai assistant to clear all the user queries and also provide valuable insights based on latest data from the vector database.

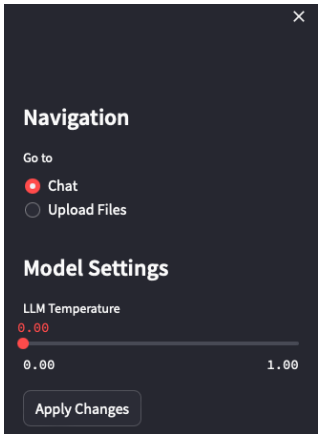

Figure 3: Response Customization tab to alter the quality of output generated by the LLM.

Through these dedicated settings to control the output, users can finetune the quality of output of the LLMs. Higher temperature values allow the LLM to be more creative with the answers generated.

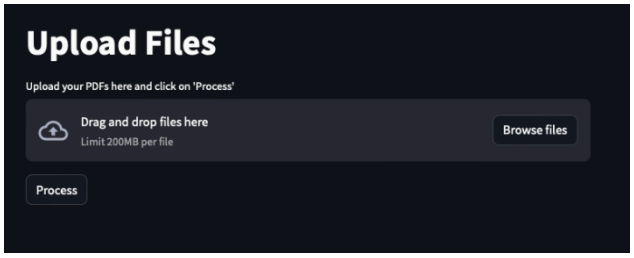

Figure 4: Upload Tab to upload custom files to the knowledge base.

This module provides user with the option to upload own data and files to the knowledge base of the Ai-Model. This further enhances the quality of output.

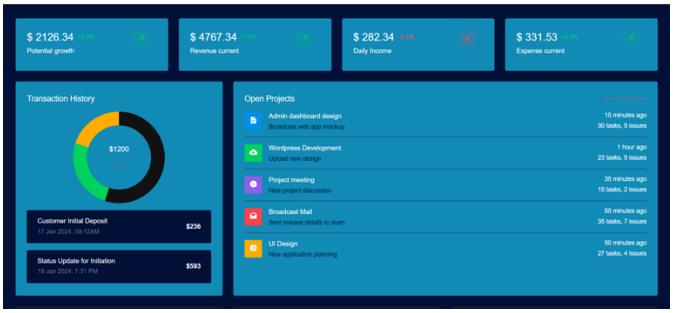

Figure 5: User Dashboard to display metrics and other data

A user-friendly UI that helps startup managers to keep track of valuable metrics like income rate, expenses and other important details like project management and transaction history.

8 Conclusion and future enhancements

Startups offer a vital avenue for addressing the pressing challenges that humanity confronts. They possess the agility to swiftly respond to issues and develop innovative solutions. The surge in the number of incubators and the growing interest of young individuals in launching their own ventures have further fueled entrepreneurship and early- stage startups in India. AI and machine learning has huge impact on the aspects of creativity and efficiency as written (Rojas & Tuomi, 2022; Oehmichen et al., 2023) . They have the potential to disrupt the entire flow of events both positively and negatively. It requires skills and proper resources to channel the development of the startup ahead.

Further Enhancements can be made in this methodology by utilizing larger and more powerful Large Language models that are trained with Billions of parameters. Since these heavier models require huge computing power which can't be obtained locally, cloud providers like Amazon Web Services provide custom AI models that have the capacity to handle huge amounts of data with ease.

Additionally, the embedding models can also be improved to handle large datasets with ease. These embedding models can be optimized to create vector embeddings with proper dimensions without any loss of information or data. The entire application can be hosted on cloud so that it is made readily available to anyone to avail the services without any issues. The application can also be scaled up and down based on demand which is very cost efficient as well.

The integration of AI and ML technologies into the initial stages of startups has the potential to revolutionize the way businesses operate. These technologies offer startups enhanced decision- making capabilities, operational efficiency, scalability, and a customer-centric approach. Moreover, they provide a competitive edge and foster innovation in research and development. As startups continue to harness the power of AI and ML, their ability to navigate the complexities of the business world is greatly amplified, as written (Vasquez & Santisteban, 2023) increasing their likelihood of success in the competitive landscape.

Authors

Varun Kumar V, Department of Computer science and Business Systems, Rajalakshmi Engineering College, 201401054@rajalakshmi.edu.in

Ruthvik S Department of Computer science and Business Systems, Rajalakshmi Engineering College, 201401039@rajalakshmi.edu.in

Mrs Jinu Sophia, Assistant Professor (SG), Department of Computer Science and Business Systems, Rajalakshmi Engineering College, jinusophia.j@rajalakshmi.edu.in

References

1. 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.
2. Cohen, S., Fehder, D. C., Hochberg, Y. V., & Murray, F. (2019). The design of startup accelerators. *Research Policy, 48*(7), 1781-1797. https://doi.org/10.1016/j.respol.2019.04.003.
3. De Cremer, D., Morini Bianzino, N., & Falk, B. (2023). How Generative AI Could Disrupt Creative Work.
4. Ghina, A., & Sinaryanti, I. (2021). The Learning Evaluation of Business Incubator's Role in Developing Technology-Based Startups at Technology Business Incubator. *The Asian Journal of Technology Management (AJTM), 14,* 35-56. https://doi.org/10.12695/ajtm.2021.14.1.3.
5. Gui, Z., Paterson, K. G., & Tang, T. (2020). Security Analysis of MongoDB Queryable Encryption. *ETH Zurich*.
6. Maradi, M. (2023). Growth of Indian start-up: A critical analysis, *17*, 180-186.
7. Marcon, A., & Ribeiro, J. L. D. (2021). How do startups manage external resources in innovation ecosystems? A resource perspective of startups' lifecycle. *Technological Forecasting and Social Change, 171,* 120965. https://doi.org/10.1016/j.techfore.2021.120965.
8. Öztürk, A. M. E. (2022). A method to improve full text search performance of MongoDB. doi: 10.5505/pajes.2021.89590.
9. Pandiaraja, P., Boopesh, K. B., Deepthi, T., Laksmi Priya, M., & Noodhan, R. (2022). An Analysis of Document Summarization for Educational Data Classification Using NLP with Machine Learning Techniques.
10. Pronin, C. B., & Ostroukh, A. V. (2023). Synthesis of Quantum Vector Databases Based on Grover's Algorithm. https://doi.org/10.48550/arXiv.2306.15295.
11. Puapongsakorn, P. L., & Brazdeikyte, E. (2023). Exploring the Integration of Artificial Intelligence in the Ideation Stage of Product Development in Swedish Startups: Challenges, Opportunities, and Tool Utilization. *Department of Business Administration*.
12. Rojas, A., & Tuomi, A. (2022). Reimagining the sustainable social development of AI for the service sector: The role of startups. *ISSN: 2633-7436*.
13. Vasquez, E., & Santisteban, J. (2023). Predicting the Success of a Startup in Information Technology Through Machine Learning. *International Journal of Information Technology and Web Engineering*. https://doi.org/10.4018/IJITWE.323657.
14. Oehmichen, J., Schult, A., & Dong, J. Q. (2023). Successfully organizing AI innovation through collaboration with startups. *MIS Quarterly Executive, 22*(1), Article 4. Available at: https://aisel.aisnet.org/misqe/vol22/iss1/4.