

A Hybrid NLP-Based Multi-Source Recommendation System for Matching Startups with Investors, Government Grants, and Incubators

Venkata Kavya Botta Dept.
of CSE (Data Science) Anil
Neerukonda Institute of
Technology and Sciences
Visakhapatnam, India

Meghana Andhavarapu
Dept. of CSE (Data Science)
Anil Neerukonda Institute of
Technology and Sciences
Visakhapatnam, India

Bharathkumar Dakara
Dept. of CSE (Data Science)
Anil Neerukonda Institute of
Technology and Sciences
Visakhapatnam, India

Abhay Udumulla
Dept. of CSE (Data Science)
Anil Neerukonda Institute of
Technology and Sciences
Visakhapatnam, India

Udhayasri Jaka
Dept. of CSE (Data Science)
Anil Neerukonda Institute of
Technology and Sciences
Visakhapatnam, India

Abstract - Startups often have trouble finding the right funding opportunities because information is spread out across different platforms, such as investor databases, government websites, and incubator listings. This paper presents an AI-powered recommendation system that employs Natural Language Processing (NLP) to link startups with suitable investors, government grants, and incubators. The system uses a hybrid approach that combines a structured rule-based scoring system with transformer-based models that measure semantic similarity. Data came from a number of places, including Startup India, OpenVC, and directories of incubators. We looked at three models: TF-IDF, BERT, and Sentence-BERT (SBERT). The experimental results show that SBERT is the best model for recommendation tasks because it does the best job of ranking based on NDCG and MAP. The system is implemented as a real-time web application using ReactJS and FastAPI, showcasing its practical utility within the startup ecosystem.

Index Terms—Startup Recommendation, SBERT, BERT, NLP, Hybrid Scoring, Information Retrieval, Venture Capital Matching

I. INTRODUCTION

Startups are very important for encouraging new ideas and helping the economy grow. However, one of the biggest problems they face is finding the right sources of funding, like venture capitalists, government grants, and incubators. These funding options are often spread out over different platforms, so startups have to manually search through, filter, and evaluate large amounts of data. Conventional recommendation systems that rely on keyword matching or manual filtering do not understand the context and often give irrelevant results. But improvements in Natural Language Processing (NLP) have led to the creation of transformer-based models like BERT, which

help computers understand the meaning of text and make more advanced recommendation systems [3].

This paper presents a hybrid NLP-based recommendation system that integrates semantic similarity with structured scoring to provide accurate and relevant matches. This system can take suggestions from many different places and is set up as a web-based application that works in real time. The main contributions of this study are a single platform that brings together investors, grants, and incubators; a hybrid scoring system that combines semantic and structured features; a comparison of the TF-IDF, BERT, and SBERT models; and a real-time web-based system that has been put into use.

II. RELATED WORK

There has been a lot of research on recommendation systems in areas like e-commerce, job matching, and academic research [1], [2]. TF-IDF and other traditional methods rely on lexical similarity [5], but they don't do a good job of capturing contextual meaning. BERT introduced transformers for contextual embeddings, improving performance on a wide range of NLP tasks [3], [9]. However, BERT is not optimized for similarity comparison tasks. Sentence-BERT (SBERT) addresses this limitation by generating sentence-level embeddings suitable for semantic similarity [4]. Hybrid recommendation systems that combine content-based filtering with rule-based constraints have shown improved performance [6]. We extend this methodology to the startup funding ecosystem.

III. DATASET AND PREPROCESSING

A. Data Sources

The system integrates data from multiple sources:

- Government Grants: 57 entries scraped from Startup India portal
 - Investors: 2546 entries from OpenVC dataset
 - Incubators: 137 entries from Startup India and Failory
- These datasets provide information such as investment thesis, industries, funding stages, geographic focus, and funding ranges.

B. Data Preprocessing

Data preprocessing involves:

- Converting text to lowercase
- Removing duplicates
- Handling missing values
- Merging multiple columns into a unified text field (combined_text)

This combined representation improves the effectiveness of semantic embedding models.

IV. METHODOLOGY

A. System Overview

The system takes the following startup inputs:

- Description
- Industry
- Stage
- Country
- Funding requirement

It outputs a ranked list of investors, incubators, and grants.

B. Text Representation Models

Three distinct models were assessed: TF-IDF [5], which serves as the baseline lexical model; BERT [3], a model utilizing contextual embeddings; and SBERT [4], which is optimized for sentence-level similarity embeddings. SBERT employs the all-MiniLM-L6-v2 model to produce its embeddings.

C. Hybrid Scoring Mechanism

The final recommendation score is computed as follows:

$$\begin{aligned} \text{Final Score} = & 0.5 \times \text{Semantic Similarity} \\ & + 0.2 \times \text{Industry Match} \\ & + 0.1 \times \text{Stage Match} \\ & + 0.1 \times \text{Geographic Match} \\ & + 0.1 \times \text{Funding Match} \end{aligned} \quad (1)$$

Semantic similarity is calculated using cosine similarity:

$$\text{Semantic Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (2)$$

Industry match is computed using Jaccard similarity:

$$\text{Industry Match} = \frac{|I_A \cap I_B|}{|I_A \cup I_B|} \quad (3)$$

Stage, geographic, and funding matches are defined as:

$$\text{Stage Match} = \begin{cases} 1 & \text{if stages match} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\text{Geographic Match} = \begin{cases} 1 & \text{if geographic locations match} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$\text{Funding Match} = \begin{cases} 1 & \text{if funding status matches} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

D. Evaluation Metrics

The system is evaluated using standard information retrieval metrics [7]:

- Precision@10
- Recall@10
- F1 Score@10
- Mean Average Precision (MAP)
- Normalized Discounted Cumulative Gain (NDCG@10)

V. SYSTEM ARCHITECTURE

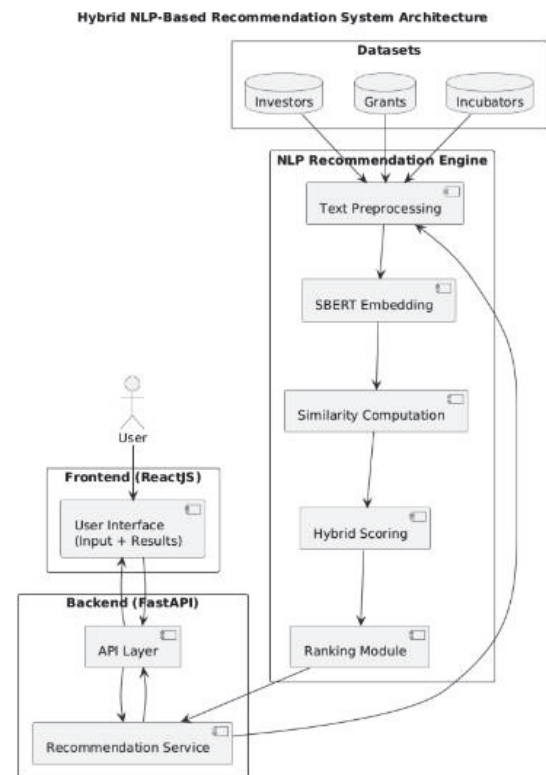


Fig. 1. System Architecture Diagram

The system consists of three main components:

- Frontend (ReactJS)
- Backend (FastAPI)
- Recommendation Engine

Workflow:

- User inputs startup details
- Data sent to backend
- Embeddings generated
- Hybrid score calculated
- Ranked recommendations returned

VI. IMPLEMENTATION AND USER INTERFACE

A. Landing Page

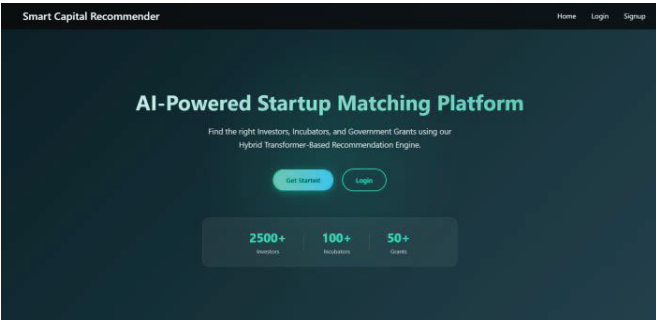


Fig. 2. Landing Page

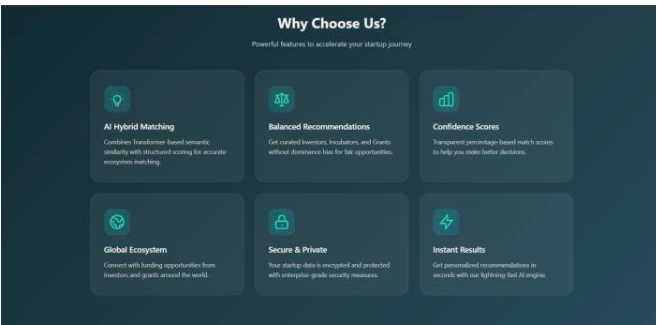


Fig. 3. Landing Page Continued

The landing page introduces the platform and highlights its features such as AI hybrid matching and confidence scores.

B. Startup Input Dashboard

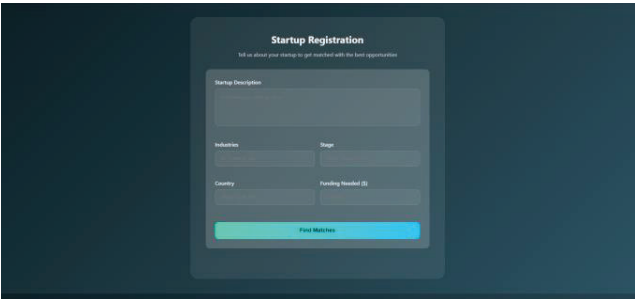


Fig. 4. Startup Input Dashboard

Users enter startup details including description, industry, stage, country, and funding requirement.

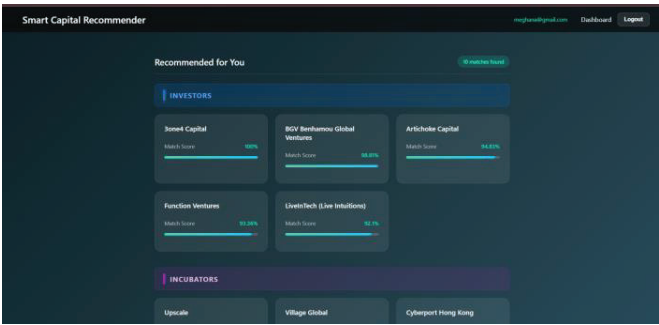


Fig. 5. Recommendation Results

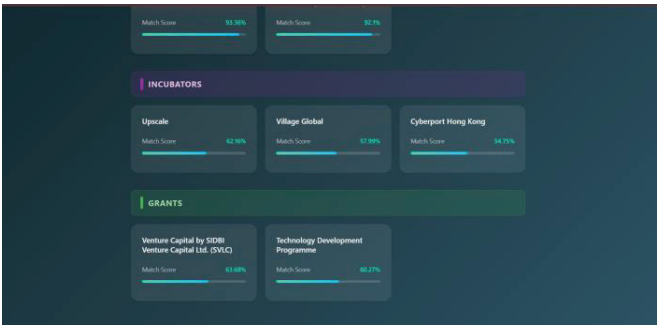


Fig. 6. Recommendation Results Continued

C. Recommendation Results

The system displays investors, incubators, government grants, and match scores with visual indicators.

VII. RESULTS AND ANALYSIS

TABLE I
MODEL PERFORMANCE COMPARISON

Model	P@10	R@10	F1@10	MAP	NDCG@10
TF-IDF	0.7	0.152	0.25	0.587	0.728
BERT	0.1	0.0217	0.0357	0.152	0.066
SBERT	0.7	0.152	0.25	0.514	0.731

Analysis:

SBERT has the highest NDCG, which means that the ranking quality is better. BERT doesn't work well because it doesn't optimize for similarity. TF-IDF does well in terms of performance, but it doesn't understand meaning. SBERT is the best choice because it balances accuracy and ranking relevance. So, SBERT is chosen as the final model.

VIII. CONCLUSION

This paper describes a hybrid NLP-based recommendation system that helps startups find funding opportunities. The system gets better recommendation accuracy and ranking quality by using structured scoring and semantic similarity together. The results show that SBERT works better than older methods and is a good fit for real-world recommendation systems. The deployed web application demonstrates the practical utility of the proposed methodology.

IX. FUTURE WORK

Future enhancements include:

- Fine-tuning SBERT for domain-specific data
- Expanding datasets with global funding sources
- Incorporating explainable AI techniques
- Adding user feedback for adaptive learning

REFERENCES

- [1] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems Handbook," Springer, 2011.
- [2] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, 2009.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL*, 2019.
- [4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP*, 2019.
- [5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, 1988.
- [6] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, 2002.
- [7] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.
- [8] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," 2013.
- [9] A. Vaswani et al., "Attention is All You Need," *NeurIPS*, 2017.
- [10] K. Sparck Jones, "A statistical interpretation of term specificity," *Journal of Documentation*, 1972.