Comprehensive Startup Success Prediction Using Textual, Structured and Network Data Models.

Rithiani M
Department of CSE (AIE)
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Chennai
m_rithani@ch.amrita.edu

Aparna S
Department of CSE (AIE)
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Chennai
aparnasureshkumar24@gmail.com

Maddineni Sravani
Department of CSE (AIE)
Amrita School of Computing
Amrita Vishwa Vidyapeetam, Chennai
sravanimaddineni24@gmail.com

SyamDev R S
Department of AI & ML
Amrita School of Computing
New Horizon College of Engineering,
Bangalore

***Abstract*— The goal of this paper is to forecast startup success by combining textual, structured and network data in a multi-modal manner. Three models, each trained on a different type of data, are used: BERT for textual data, FFNN for structured data and GNN for network data. 99.15% accuracy is attained by BERT with a low loss of 0.03, 93.82% accuracy is attained by FFNN with a loss of 0.16 and 70% accuracy is attained by GNN with a greater loss of 0.44. Using a stacking ensemble technique the outputs from all three models are merged and the predictions are integrated by a meta-model to improve the overall accuracy. By utilising the advantages of each model individually, this ensemble approach enhances the ability to predict company success and offers insightful information to all parties involved in the startup ecosystem.**

Keyword – Ensemble Learning, BERT, FFNN, GNN, Stacking Model, CrunchBase Dataset, GCN.

I. INTRODUCTION

Research done so far in this fast-emerging domain of startup success prediction has been largely dominated by applying particular machine learning models or applying certain categories of data in order to predict business outcomes. Most of these efforts miss the complexities involved in startup ecosystems but have laid down basic bases in predictive analytics in this field. Further, studies like Deller man et al. (1) and Xi et al. (2) deal with structured data such as financial indicators and founder experience but do not take into account the critical influence of network dynamics or the richer information contained in textual data. Our study, therefore, incorporates three types of data: structured financial data, textual data, and network-based data into a hybrid ensemble model in order to tackle this complex issue comprehensively and integrated. The textual data - which can carry sequential and nuanced information from analyses that include the market, of the founders, as well as business descriptions- is processed by this model using BERT. This makes our approach unique from the previous studies that utilized only financial data, like Sadatrasoul et al. (4), that could not capture deeper content in the text-based sources of data. Among these structured financial data are metrics including funding rounds, industry and years of operation to which the second component-feedforward neural network (FFNN) is applied. This is a strength of our model, as we do not compare with those methods like Pan et al. (5), which did not take into account the deeper relational patterns between companies and instead focused on rounds related to finance.

The integration of GNNs, to predict links between companies based on criteria like shared geography, similarities in funding, and industrial linkages, really gets our concept to stand out. An important aspect often left behind in previous models, like Singh and Singh (7), which focuses on the individual attributes without contemplating how startups influence and interact with each other in a networked environment, the GNN component captures the effects that occur within the startup ecosystem. They referred to social networks in their paper by Liang and Yuan (8) but did not apply those observations to any kind of predictive framework. This model performance strategy can take on typical problems such as data imbalance, a significant factor in the prediction of the success of a startup; hence this approach like SMOTE is used in order for our model to give an even representation of the successful and unsuccessful firms. In this case, it solves one of the reoccurring difficulties in the previous studies like Ross et al. (6), with their imbalanced datasets that affect the model's accuracy. We also use cross-validation to ensure that our model generalizes well, which reduces overfitting and increases its applicability in all startup ecosystems-this is one limitation of models like Ferrati (9). Our model, with the three sources of data and the application of complex machine learning techniques, provides a more robust prediction framework, much more reliable and exhaustive as well as accurate. It evaluates the specific characteristics of each endeavor and the complex relationship network that make up the startup ecosystem. Using this multi-dimensional perspective, we may provide a more precise and relevant prediction to investors, entrepreneurs, and other stakeholders. That makes our project stand out compared with most other models, and we bring sophistication to the study of company success predictions.

II. LITERATURE SURVEY

In order to increase the accuracy of startup success prediction the research [1] "Finding the Unicorn: Predicting earlystage Startup Success Through Hybrid Intelligence" presents a methodology that combines human expertise with machine learning. CrunchBase data on investment size, team

experience and market dynamics are used in the model. Using data from CrunchBase and TechCrunch the paper [2]"Forecasting M&A Success Using Machine Learning" forecasts M&A and IPO results with an emphasis on investment details, founder experience, and location. Data imbalance is addressed with techniques like SMOTE. LSTMs and CNNs are reviewed for business success prediction using CrunchBase and Kaggle data in the research [3] "Modelling and Prediction of Business Success Using Deep Learning," emphasising their capacity to capture intricate correlations between founder qualities and market performance. The study [4] "Hybrid Business Success Versus Failure Classification Prediction Model" shows enhanced accuracy for early-stage ventures in predicting startup success in the Iranian ecosystem through the combination of decision trees and logistic regression using financial and HR data. The study [5] "CrunchBase-Based Success Prediction Using Logistic Regression" highlights the importance of early financial measures and fundraising rounds for long-term success by predicting company success using logistic regression, k-nearest neighbours and random forests.

In order to focus on sustainability scores derived from financial data, market conditions and founder experience the study [6] "Startup Sustainability Forecasting with AI" suggests a decision support system that uses random forests and natural language processing. Using neural networks on financial, team composition and operational data the research "A Novel Approach for Predicting Startup Success Using Deep Learning Algorithms" [7] demonstrates that team size and funding stage are crucial to long-term success. The study [8] "Social Network-Based Prediction of Startup Financing Behaviour" examines the relationships between founders, investors and success rates in order to forecast investor behaviour and financing results using machine learning and data from CrunchBase. In order to anticipate funding success and market performance, the study [9] "Entrepreneurial Finance: Emerging Approaches Using Machine Learning" reviews machine learning approaches in the field of entrepreneurial finance. It does this by analysing data from startup ecosystems and financial records. SVM and neural networks are used in the paper [10] "Machine Learning Approach Towards Startup Success Prediction" to forecast business performance using high-dimensional time series data, including specific financial and market parameters.

In order to forecast investment behaviour and identify important success factors the study [11] "Investment Behaviour Prediction Using Machine Learning" employs logistic regression and gradient boosting on market, financial and geographic data. In the study [12] "Predicting Startup Success with Capital Investment Data," the relationship between investment amount, market conditions and founder traits is modelled for the purpose of predicting company success using CrunchBase data and machine learning techniques such as random forests. Using financial and performance data the study [13] "Assessing the Impact of Accelerators on Startup Success" examines accelerator programmes across national boundaries and concludes that accelerator support greatly improves long-term success using machine learning models. The study [14] "Factors Affecting Startup Performance" examines the competitiveness of startups through the use of big data and machine learning, examining variables such as investment quality and market dynamics and making use of random forests for prediction. The study [15] "Stock Market Dynamics in Startup Success Prediction" investigates how stock market behaviour such as price volatility and fundraising stages predicts company success using machine learning methods like logistic regression and SVM.

III. METHODOLOGY

This paper applies a hybrid ensemble learning model which blends three different sources of data: textual, network-based, and structured financial data to predict the probability of startup success. Such a multifaceted method ensures that all important aspects of startup success are fully recorded. The proposed approach makes use of three specific machine learning techniques: Graph Neural Networks (GNN) for relational data, Feedforward Neural Networks (FFNN) for structured financial data, and BERT for textual data processing. Once these three models are stacked together, the following more elaborate prediction framework, accounting for various dimensions of a startup's profile, is created.

A. Data Collection And Processing

The dataset that was relied upon in this paper was the CrunchBase dataset. One of the prominent sources that provide comprehensive data on companies, from textual descriptions to network ties and financial success, this is indeed a state-of-the-art resource. Data were categorized into three categories: structured, textual, and network-based data. Each represented a different aspect of operational features of a startup.

B. Data Types

Structured Financial Data: These are quantitative measures that can provide information about the firm's performance and prospects, such as its investment amount, number of funding rounds, years of operation, equity crowdfunding, or other relevant financial metrics.

Textual Data: There were the use of descriptive text data, including market classifications, team composition and firm descriptions. These textual inputs richly provide qualitative insights into these descriptive data, capturing startup traits critical to success prediction but quite often difficult to measure.

Network Data: Links between startups were recorded, including mutual funders, industry contacts, and similar proximity. This data gives important information to understand how startups are affected by their environment and would be required in modelling the connections between startups.

C. Data PreProcessing

All types of data underwent particular preparation processes to ensure the quality and compatibility of data in modelling. Structured Data Processing Scaling: To normalize the input data and eliminate biases caused by differing feature magnitude, numerical features such as total funding and number of financing rounds were scaled by accepted scaling techniques.

Handling Missing Data: Since one would consider this type of data to be important and also self-consistent, missing values in the structured data were imputed by proper

techniques, such as median imputation for numerical fields and mode imputation for categorical data. Textual data processing: Tokenization and padding: This is through BERT's tokenizer that chops the text into tokens and translates this into numeric embeddings; the steps elaborated above were applied to the textual data. Since the sequence of all texts have to be of equal length, to make the model intake data faster in digestion, a padding component was added.

Contextual Embeddings: Use of BERT on the startup descriptions for the conversion into contextual embeddings. This approach gives deep insights into the qualitative features of the businesses by capturing the meaning and context of words in their particular use. Network Data Preprocessing Node Mapping and Edge Creation: Companies in the dataset were mapped to nodes in the network with edges created between nodes depending on common characteristics such as funding rounds, industry, and location. This type of network structure helped GNNs learn links and relationships about the startups.

To identify the node pairs in the network graph, which provide an input to the GNN sharing edges, an edge index tensor was constructed.

D. Algorithm Implementation

After preprocessing the data three different machine learning techniques, one for each category of data, were used to build the prediction model. These models were selected to handle textual, structured, and network data, respectively: BERT, Feedforward Neural Networks (FFNN), and Graph Neural Networks (GNN). The foundation of the framework for predicting startup success was created by combining the outputs of several models using a stacking ensemble technique.

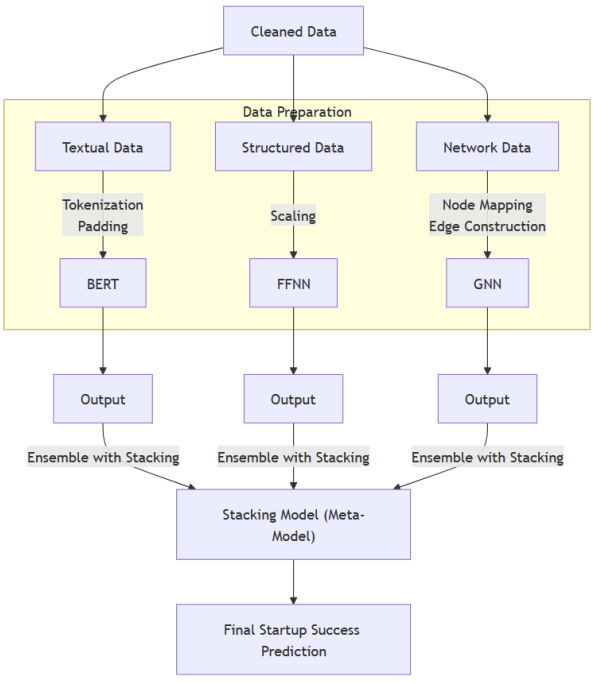

Fig. 1. Architecture of the system

1. BERT (Bidirectional Encoder Representations from Transformers): BERT is an extremely efficient transformer-based architecture for problems involving natural language processing. BERT was used in this research to process textual data, including market classifications, startup firm profiles and other qualitative information. Tokenization: BERT's tokenizer, which divides text into smaller pieces known as tokens, was used to tokenize the incoming textual data. After that, these tokens are represented numerically. Embedding: Every token is associated with a high-dimensional embedding that encompasses contextual information in addition to semantic meaning. The bi-directional attention mechanism of BERT makes sure that the context is recorded from both the left and the right. levels and Attention: To understand links between tokens across several transformer levels 12 layers in our case the token embeddings are subjected to self-attention.
2. Feedforward Neural Network (FFNN) for Structured Data: A Feedforward Neural Network was used to handle structured financial data. A common deep learning architecture that works well with structured inputs like numerical data points is called FFNN. Input Layer: The input layer receives the structured financial data, such as industry categories, funding quantities, and equity crowdfunding. Dense Layers: There are three completely connected layers (Dense Layers) in the FFNN. There are 128 neurones in the first dense layer, 64 neurones in the second, and 32 neurones in the third. The ReLU activation function is used in each of these layers to add non-linearity. Output Layer: Using structured data, a softmax or sigmoid activation (for binary classification) is applied to produce the final output, which forecasts the chance of startup success.
3. Graph Neural Network (GNN) for Network DataThe network data which shows relationships between businesses based on shared characteristics including fundraising rounds, geographic proximity, and common investors, was processed using the Graph Neural Network (GNN). Node and Edge Representation: In the graph, every startup is a node, and edges are formed between nodes according to common attributes. For instance, there is a connection between startups in the same industry or with comparable funding histories. Graph Convolutional Layers: Three Graph Convolutional Network (GCN) layers make up the GNN architecture. The initial layer of the GCN evaluates each node's input features (geography, funding amounts, etc.) and uses convolutional techniques to compile data from nearby nodes. To iteratively improve the node representations, this procedure is repeated in later GCN layers. Activation and Output: To classify the startups according to their likelihood of success, a softmax activation is given to the final output after each GCN layer, followed by a ReLU activation.

Ensemble Learning with Stacking: Utilising stacking for group learning is the last stage of the model creation process. Based on the different types of data that they each handle, the three models - FFNN, GNN and BERT produce different results. The stacking model a meta-model receives these outputs and uses the predictions from all of the models to arrive at a conclusion. Meta-Model: Taking the outputs from BERT, FFNN and GNN the meta-model is a straightforward classifier (such as logistic regression or

another neural network) that produces the final binary prediction (successful/unsuccessful startup). The Reason for Stacking: By stacking, the model can take advantage of each underlying model's advantages, producing a prediction that is more thorough and precise. Textual information is captured by BERT, financial data is learned by FFNN, and the relationship impacts between startups. The model can produce extremely accurate predictions because of this combination of specialised models, each of which handles a different sort of data. This allows the model to provide insights into the myriad elements that contribute to startup success. The model's performance is further improved by using ensemble learning through stacking, which makes sure that no one data type controls the prediction process.

IV. RESULTS AND DISCUSSION

In this research work, we had employed three different types of data- structured data, textual data, and network data respectively, so that we developed three models: Feedforward Neural Network (FFNN), BERT, and Graph Neural Network (GNN) to predict the success of the startups. For all the performance evaluation of the models, it substantially focused on the key performance indicators such as accuracy and loss.

S. No	Model	Accuracy	Loss
1	FFNN	93.82 %	0.16
2	BERT	99.15 %	0.03
3	GNN	70.00 %	0.44

Fig. 2. Accuracy and Loss of three Models

As seen above, clearly, BERT performs way better than the rest of the models as it achieves an accuracy of 99.15% with the least possible loss at 0.03, which reflects how skilled it is in picking up textual information. Structured FFNN, being trained on financial data, was also performing quite well with a low loss of 0.16 and 93.82% accuracy. GNN, simulating network relationships among the startups, faced a relatively higher loss of 0.44 with relatively lower accuracy at about 70%.

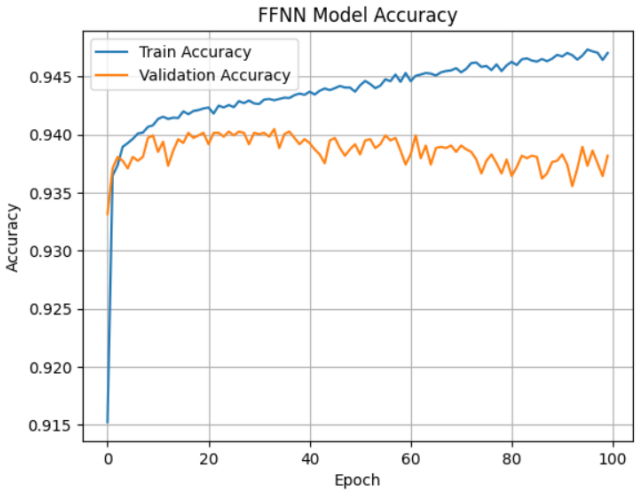

Fig. 3. Accuracy plot for FFNN

As can be observed, BERT performs noticeably better than the other models, attaining an accuracy of 99.15% with the least amount of loss (0.03), demonstrating how well it can capture textual information. Trained on organised financial data, FFNN also performed well, with a low loss of 0.16 and 93.82% accuracy. On the other hand, GNN which simulates the network relationships across startups, had a larger loss of 0.44 and a comparatively lower accuracy of 70%.

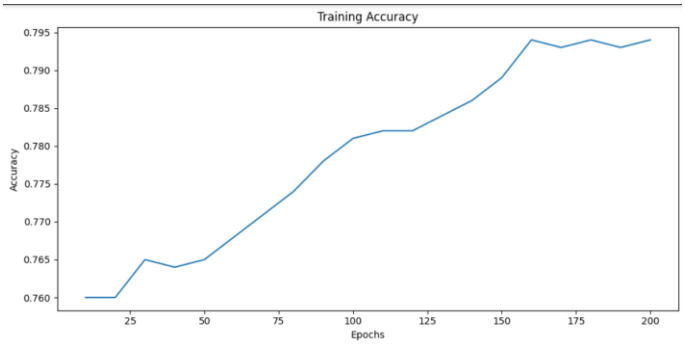

Fig. 4. Accuracy Plot for GNN

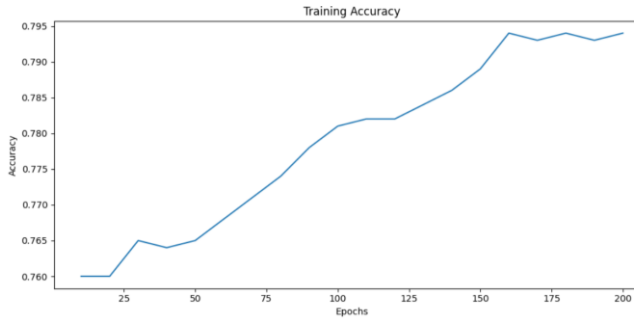

Fig. 5. Accuracy plot for BERT

Ensemble Performance: We used a stacking ensemble technique, merging the predictions from FFNN, BERT, and GNN to form a meta-model, in order to take advantage of each model's advantages. Because each model was able to offer its unique insights FFNN for financial indicators, GNN for network interactions, and BERT for textual data the ensemble technique increased the overall prediction accuracy. Accuracy and Loss of Training The training accuracy and training loss curves for the models over ten epochs are displayed in the following figures: Training Loss: An optimal learning process is shown by the ensemble model's training loss, which as Figure 1 illustrates gradually decreasing across the epochs.

V. CONCLUSION

The results of this investigation validate the significance of utilising a multi-model ensemble methodology. Through the integration of textual, structured, and network data, we offer a more all-encompassing perspective on the elements that lead to startup success. By combining the unique strengths of BERT, FFNN and GNN the ensemble model performs better than standard models attaining higher prediction accuracy and better managing the intricacies of startup ecosystems.

REFERENCES

[1] Dellermann, Dominik, Nikolaus Lipusch, Philipp Ebel, Karl Michael Popp, and Jan Marco Leimeister. "Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method." *arXiv preprint arXiv:2105.03360* (2021).

[2] Karatas, Tugce, and Ali Hirsa. "Predicting Status of Pre and Post M&A Deals Using Machine Learning and Deep Learning Techniques." *arXiv preprint arXiv:2110.09315* (2021).

[3] Gangwani, Divya, and Xingquan Zhu. "Modeling and prediction of business success: a survey." *Artificial Intelligence Review* 57.2 (2024): 44.

[4] Sadatrasoul, Seyed M., O. Ebadati, and R. Saedi. "A hybrid business success versus failure classification prediction model: A case of iranian accelerated start-ups." *Journal of AI and Data Mining* 8.2 (2020): 279-287.

[5] Maarouf, Abdurahman, Stefan Feuerriegel, and Nicolas Pröllochs. "A fused large language model for predicting startup success." *European Journal of Operational Research* (2024). M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Jorzik, Philip, et al. "Sowing the seeds for sustainability: A business model innovation perspective on artificial intelligence in green technology startups." *Technological forecasting and social change* 208 (2024): 123653.

[7] Panduri, Bharathi, et al. "An efficient and sustainable novel approach for prediction of start-up company success rates through sustainable machine learning paradigms." *E3S Web of Conferences*. Vol. 430. EDP Sciences, 2023

[8] Liang, Yuxian Eugene, and Soe-Tsyr Daphne Yuan. "Predicting investor funding behavior using crunchbase social network features." *Internet Research* 26.1 (2016): 74-100.

[9] Liang, Yuxian Eugene, and Soe-Tsyr Daphne Yuan. "Predicting investor funding behavior using crunchbase social network features." *Internet Research* 26.1 (2016): 74-100.

[10] Ünal, Cemre, and Ioana Ceasu. *A machine learning approach towards startup success prediction*. No. 2019-022. IRTG 1792 Discussion Paper, 2019.

[11] Silva, Thiago Christiano, Benjamin Miranda Tabak, and Idamar Magalhães Ferreira. "Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies."

[12] Krishna, Amar, Ankit Agrawal, and Alok Choudhary. "Predicting the outcome of startups: less failure, more success." *2016 IEEE 16th international conference on data mining workshops (ICDMW)*. IEEE, 2016.

[13] Polo García-Ochoa, Celia, Carmen De Pablos Heredero, and Francisco José Blanco Jiménez. "How business accelerators impact startup's performance: Empirical insights from the dynamic capabilities approach." *Intangible Capital* 16.3 (2020): 107-125.

[14] Kim, Jongwoo, Hongil Kim, and Youngjung Geum. "How to succeed in the market? predicting startup success using a machine learning approach." *Technological Forecasting and Social Change* 193 (2023): 122614.

[15] Jayanth, K. Krishna, et al. "Intent Recognition Leveraging XLM-RoBERTa for Effective NLU." *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 2024.

[16] Krishnan, Rohith, et al. "Enhancing Brain Tumor Diagnosis: A CNN-Based Multi-Class Classification Approach." *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. Vol. 2. IEEE, 2024.

[17] R. Venkatakrishnan, M. Rithani, G. Bharathi Mohan, V. Sulochana and R. PrasannaKumar, "Revolutionizing Talent Acquisition: A Comparative Study of Large Language Models in Resume Classification," *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, 2024, pp. 1-6, doi: 10.1109/ICITIIT61487.2024.10580109.

[18] Kodavati, T., Rithani, M., Venkatraman, K., & SyamDev, R. S. (2023, December). Detection and Classification of Arrhythmia Using Hybrid Deep Learning Model. In *2023 International Conference on Next Generation Electronics (NEleX)* (pp. 1-6). IEEE.

[19] Ajith, Sangeeth, M. Rithani, and R. S. SyamDev. "Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach." *2023 Seventh International Conference on Image Information Processing (ICIIP)*. IEEE, 2023.

[20] Reddy, Keshavagari Smithin, et al. "A Comparative Analysis: Enhancing Baby Cry Detection with Hybrid Deep Learning Techniques." *2023 International Conference on Next Generation Electronics (NEleX)*. IEEE, 2023.