

Natural language processing on early-stage companies’ description to predict future success

Chun Hin, Ma

Physics Student, University of Oxford, England

December 26, 2022

Abstract

Venture capital(VC) industry invests in early-stage companies that are believed to have long-term growth potential. Machine learning come to aid to support VCs make investment decisions. This study provides an NLP model to predict outcome of 5843 early-stage companies using only their self description as input features. The outcome of each startup is either 0 (unsuccessful) and 1 (successful). The data set labelled 2156 entries of successful and 3687 entries of unsuccessful startups respectively. The model returns an accuracy of 89%, a precision of 89% and a recall of 80%, which is performing about 50% better in performance than the baseline model of random classification with 49% accuracy, 37% precision and 51% recall.

1 Introduction

Investing in early-stage companies can often be risky on unproven companies and thus requires careful examinations on every information that is available to the VCs. One could utilize platforms like *techcrunch* or *crunchbase* to scout potential companies that may align with specific VCs’ investment strategies and interests. New and young companies often turn to VCs for initial funding to scale their business in exchange for the companies’ own equities. If the business manage to grow, the rewards for the VCs could be substantial. However, it has been estimated 3 out of 4 venture-backed startups fail [1], VCs have to bear in mind how to avoid unsuccessful investment due to the high failure rate of startups. While each VC has their unique portfolio strategies, it is only reasonable one would leverage the rapidly emerging machine learning methods to assist in data analysis and manage portfolio [2].

In general, machine learning can support VCs in various ways, including but not limited to i) optimising portfolio management, ii) identifying potential business partners and competitors [3], [4], iii) creating startup valuation models [5], [6], iv) creating recommendation systems to match startups with VCs [7]. Machine learning methods provide data-driven non-obvious insights for investment decisions, which may remind investors of potential continuation bias towards decisions made [8]–[11]. Machine learning has also been applied on predicting the behaviours of various VCs, where attempts have been made to classify whether a startup would be invested by a certain group of VCs [12]–[14].

In this paper, the author focus on leveraging machine learning classification algorithms to predict the future of a startup into 1 (successful) and 0 (unsuccessful). There are a variety of approaches on this prediction method [15]–[20]. For example, by quantizing founders’ background for classification purposes, one could achieve a precision of about 0.8 [21], [22]. It has also been found that a few characteristics of founders such as team size, academic abilities and previous positions at different companies do have some predictive power for future outcomes [23]. Employing natural language processing methods via BERT transfer learning or Word2Vec could also be seen to study startups [24]–[26].

2 Methods

For this study, companies that have more than USD \$500M valuation either through an IPO (initial public offering), M&A (merger and acquisition) or large funding round (more than \$150M funding) are labelled as successful (1), whereas companies that raised more than \$4M but less than \$10M which

were founded between 2010 and 2016 are labelled as unsuccessful (0). The argument is that they were unable to move as fast as their successful counterparts. The data set labelled 2156 entries of successful and 3687 entries of unsuccessful startups respectively, making up a total of 5843 entries.

2.1 Data preprocessing

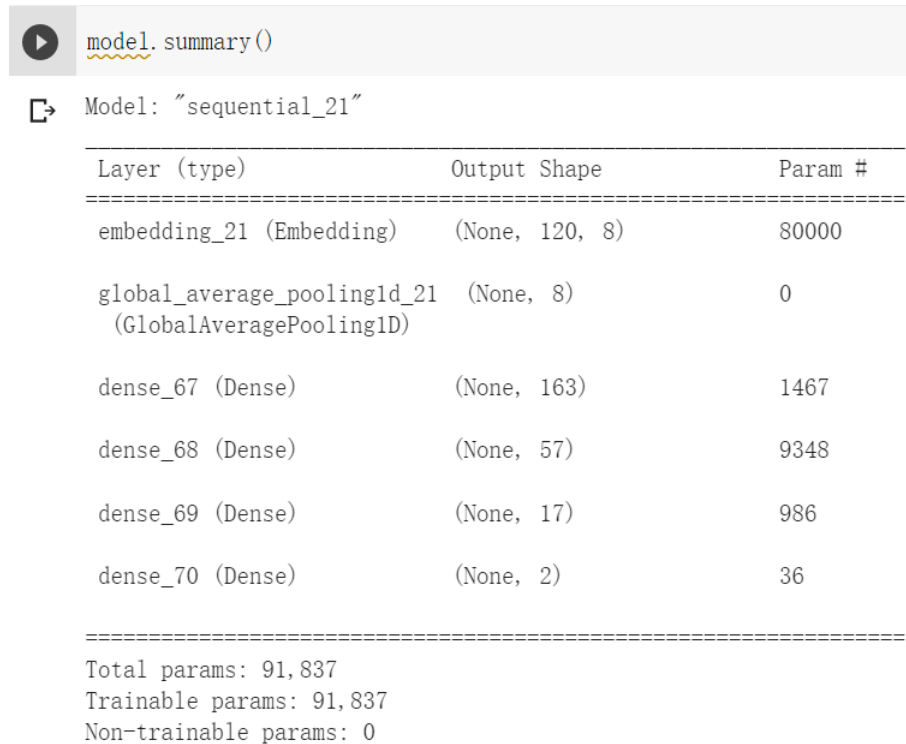
In this approach, BERT will not be used. the author argues BERT is over-killing this problem since it has been trained on BooksCorpus and English Wikipedia, which do not capture much relations between words in the business industry, but more in the academic and linguistic sense.

The only feature used for the classification algorithm is the description of the companies. It is passed to a function that remove stopwords like *am*, *at*, *about* from the description that do not carry much meaning in the texts. The first word of the name of the companies is also considered as a stopword. While most companies have the second or third word of the their names capturing the essence of their nature, a few of them has the first word capturing the essence. One could argue the performance of this approach could be improved by cherry-picking the stopwords.

The successful and unsuccessful data is then shuffled together and 70% of them are split into training, the remaining 30% are used as validation. The vocabulary in the training sentences are tokenized using the tensorflow keras package. OOV token is used to replace out-of-vocabulary words. The dimension of the dense embedding layer of the model is chosen to be 8. All description of the startups are padded into 85 words with post padding strategy.

2.2 Model

The neural consists of an embedding layer, followed by 'GlobalAveragePooling1D' layer, a dense layer with 163 neurons using softmax activation, a dense layer with 57 neurons using sigmoid activation, an ordinary dense layer with 17 neurons, and at last a dense layer with 2 neurons using softmax activation. The model is compiled with the "sparse categorical crossentropy" loss function, and an 'Adam' optimizer.



```
model.summary()
```

Model: "sequential_21"

Layer (type)	Output Shape	Param #
embedding_21 (Embedding)	(None, 120, 8)	80000
global_average_pooling1d_21 (GlobalAveragePooling1D)	(None, 8)	0
dense_67 (Dense)	(None, 163)	1467
dense_68 (Dense)	(None, 57)	9348
dense_69 (Dense)	(None, 17)	986
dense_70 (Dense)	(None, 2)	36

=====
Total params: 91,837
Trainable params: 91,837
Non-trainable params: 0

Figure 1: A summary of the neural network model used.

3 Result

The model is trained with 20 epochs and returns an accuracy of 89%, a precision of 89% and a recall of 80%. Figure 2 shows the confusion matrix of the result. Comparing with a zero rate classifier as a baseline, it will classify all startups into 0, and thus a baseline accuracy of 62%. For comparison, a baseline model of random classification is also implemented, the model returns an accuracy of 49%, a precision of 37% and a recall of 51%. Figure 3 shows the confusion matrix of such model. The accuracy, precision and recall of NLP model is performing about 50% better than the baseline model, indicating the description of the companies have a predictive power over their possibility of future success. Figure 4 shows how the training accuracy and the validation accuracy changes with respective to the number of epochs.

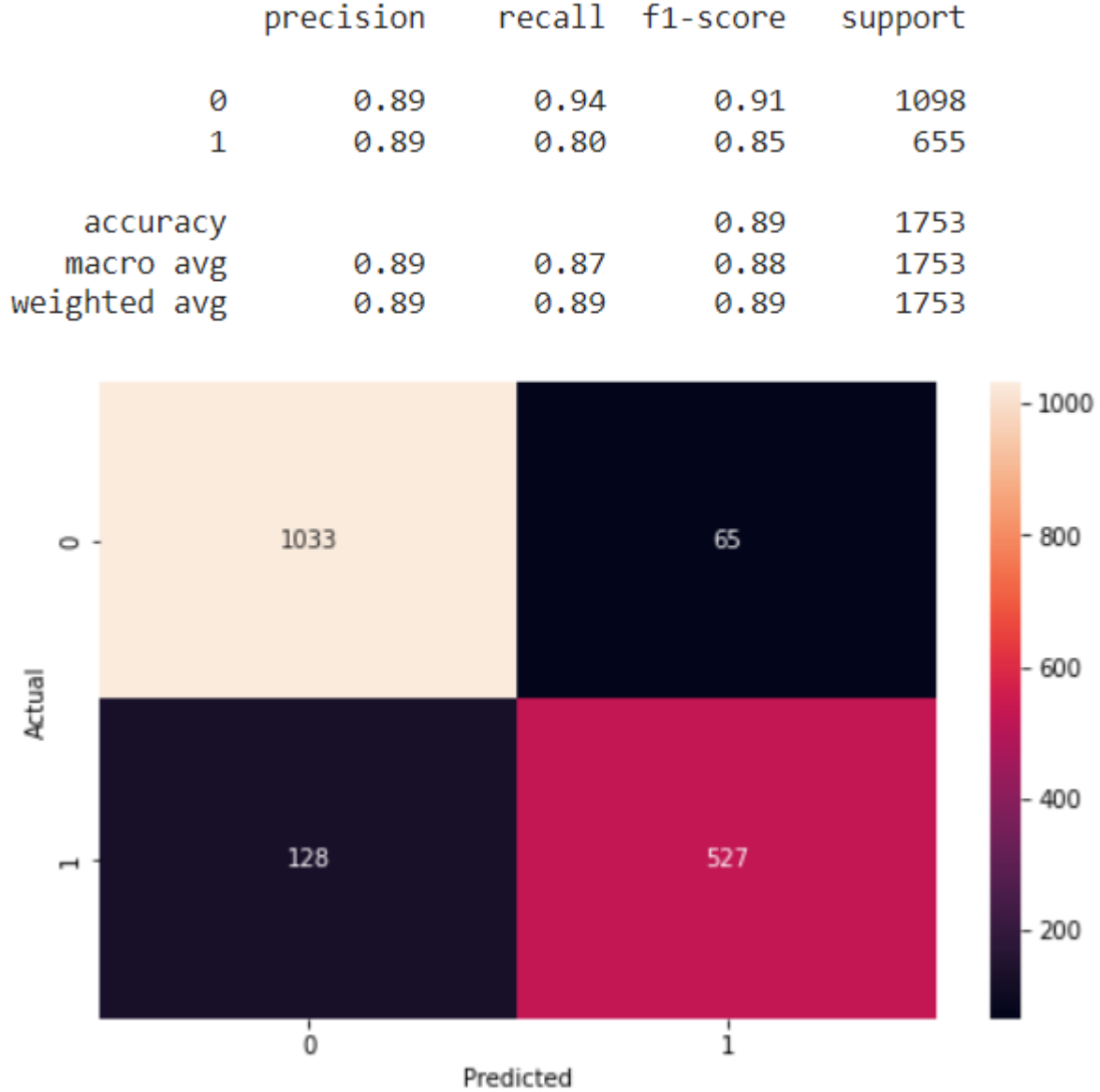


Figure 2: The confusion matrix heatmap of the result

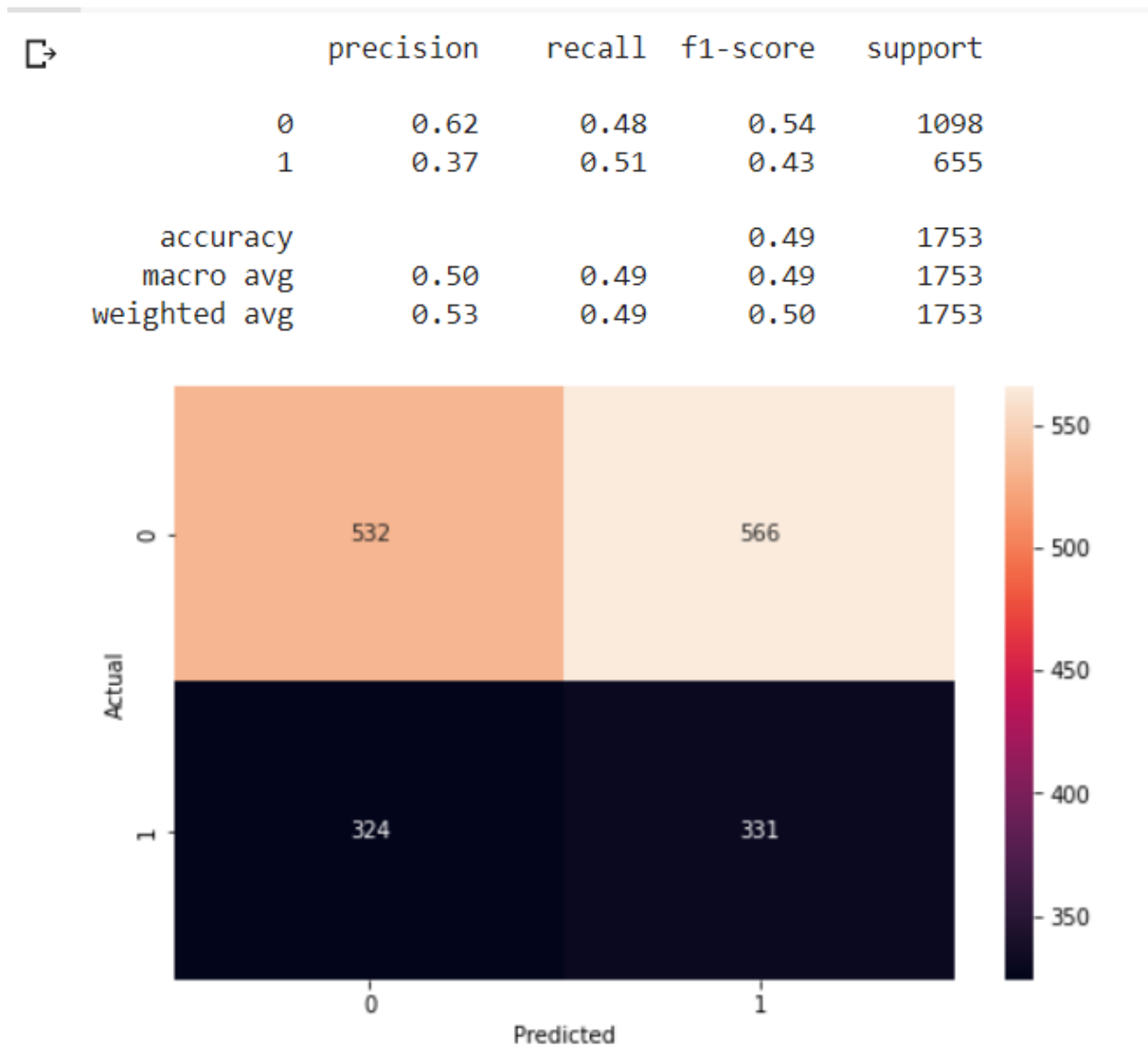


Figure 3: The confusion matrix heatmap of the baseline random classification model

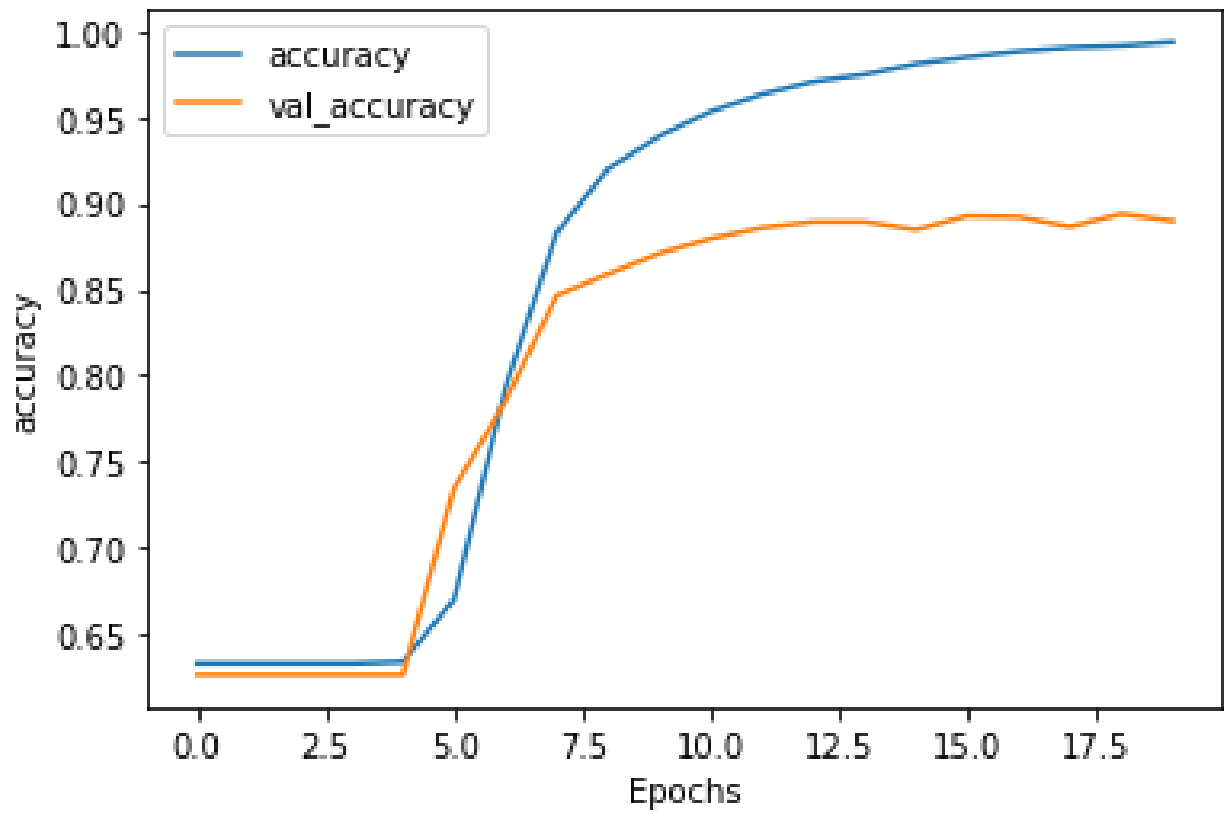


Figure 4: Training accuracy and the validation accuracy of the NLP model with respect to the number of epochs.

4 Conclusion

The author concludes the self description of the companies have an indicative power on their future success. The accuracy is 2 times the baseline accuracy of a random classification. However, it has been shown the founder’s background also have a predictive power over the outcome. One could improve this result by combining two via a voting system, or cherry picking the stopwords, fine tuning the hyperparameters. One could also employ genetic neural network to search for models with higher accuracy with an appropriate fitness function [27]–[29]. A lot of future work could be done, the author just could not encapsulate everything in a single paper.

5 Acknowledgements

The author would like to thank Yigit Ihlamur (co-founder of Vela Partner) for providing the data of the successful and unsuccessful startups and his invaluable guidance and feedback on this work.

References

- [1] D. Gage, “The venture capital secret: 3 out of 4 start-ups fail,” *The Wall Street Journal*, 2012. [Online]. Available: <https://www.wsj.com/articles/SB10000872396390443720204578004980476429190>.
- [2] F. Corea, “Artificial intelligence and venture capital,” 2018. [Online]. Available: <https://francesco-ai.medium.com/artificial-intelligence-and-venture-capital-af5ada4003b1>.
- [3] I. Sahbazoglu and Y. Ihlamur, “Project weave: Startup knowledge graph,” *Vela Partners*, 2022.
- [4] J. Piskorz and Y. Ihlamur, “Midas touch: Investor similarity scoring,” *Vela Partners*, 2022.
- [5] C. Caruso, F. Enriquez, A. Oshotse, and G. Pradeep, “Algorithmic venture capital,” 2020. [Online]. Available: http://cs230.stanford.edu/projects_fall_2020/reports/55791766.pdf.
- [6] M. Garkavenko, H. Mirisaei, E. Gaussier, A. Guerraz, and C. Lagnier, *Valuation of Startups: A Machine Learning Perspective*, D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, Eds. Springer International Publishing, 2021, ISBN: 978-3-030-72113-8. DOI: [10.1007/978-3-030-72113-8_12](https://doi.org/10.1007/978-3-030-72113-8_12).
- [7] S. Xu, Q. Zhang, L. Lü, and M. S. Mariani, “Recommending investors for new startups by integrating network diffusion and investors’ domain preference,” *Information Sciences*, vol. 515, pp. 103–115, 2020, ISSN: 0020-0255. DOI: <https://doi.org/10.1016/j.ins.2019.11.045>.
- [8] R. Bueschen, “The surprising bias of venture capital decision-making,” 2015. [Online]. Available: <https://techcrunch.com/2015/09/24/the-surprising-bias-of-venture-capital-decision-making/>.
- [9] D. Khanin and R. Mahto, “Do venture capitalists have a continuation bias?” *Journal of Entrepreneurship*, vol. 22, pp. 203–222, Sep. 2013. DOI: [10.1177/0971355713490818](https://doi.org/10.1177/0971355713490818).
- [10] T. Hull, “The effect of venture capitalists straying from their industry comfort zones,” *2nd Emerging Trends in Entrepreneurial Finance Conference*, 2018. DOI: <http://dx.doi.org/10.2139/ssrn.3163955>.
- [11] D. Davenport, “Predictably bad investments: Evidence from venture capitalists,” 2022. DOI: <http://dx.doi.org/10.2139/ssrn.4135861>.
- [12] T. C. Ho and Y. Ihlamur, “Project proxy – predicting success of early stage startups,” *Vela Partners*, 2022.
- [13] R. Zhao and Y. Ihlamur, “Project proxy - predicting startup success,” *Vela Partners*, 2022.
- [14] H. Wang and Y. Ihlamur, “Predicting decisions by successful investors using machine learning,” *Vela Partners*, 2022.
- [15] T. Hengstberger, “Increasing venture capital investment success rates through machine learning,” *Imperial College London (master thesis)*, 2020. [Online]. Available: https://www.imperial.ac.uk/media/imperial-college/faculty-of-natural-sciences/departments-of-mathematics/math-finance/HENGSTBERGER_THOMAS_01822754.pdf.

- [16] J. Arroyo, F. Corea, G. Jimenez-Diaz, and J. A. Recio-Garcia, "Assessment of machine learning performance for decision support in venture capital investments," *IEEE Access*, vol. 7, pp. 124 233–124 243, 2019. DOI: [10.1109/ACCESS.2019.2938659](https://doi.org/10.1109/ACCESS.2019.2938659).
- [17] G. Ross, S. Das, D. Sciro, and H. Raza, "Capitalvx: A machine learning model for startup selection and exit prediction," *The Journal of Finance and Data Science*, vol. 7, pp. 94–114, 2021, ISSN: 2405-9188. DOI: <https://doi.org/10.1016/j.jfds.2021.04.001>.
- [18] S. Varma, "Machine learning based outcome prediction of new ventures: A review," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU*, vol. 9, 2021. DOI: <https://doi.org/10.1016/j.jfds.2021.04.001>.
- [19] L. Cao, V. von Ehrenheim, S. Krakowski, X. Li, and A. Lutz, *Using deep learning to find the next unicorn: A practical synthesis*, 2022. DOI: [10.48550/ARXIV.2210.14195](https://doi.org/10.48550/ARXIV.2210.14195).
- [20] J. Rubruck and Y. Ihlamur, "The use of temporal features to predict startup success," *Vela Partners*, 2022.
- [21] T. E. J. Moxham and Y. Ihlamur, "Evaluation of early stage startup success using investor rankings & founders background," *Vela Partners*, 2022.
- [22] A. A. F. M. Petronilia and Y. Ihlamur, "Start-up success prediction tool," *Vela Partners*, 2021.
- [23] P. Gompers, A. Kovner, J. Lerner, and D. Scharfstein, "Performance persistence in entrepreneurship," *Journal of Financial Economics*, vol. 96, no. 1, pp. 18–32, 2010, ISSN: 0304-405. DOI: <https://doi.org/10.1016/j.jfineco.2009.11.001>.
- [24] L. Secilmis and Y. Ihlamur, "Maverick: Guiding venture capital investment through deep learning & nlp," *Vela Partners*, 2022.
- [25] I. Cheung and Y. Ihlamur, "Sbert based sentence similarity search," *Vela Partners*, 2022.
- [26] L. Song and Y. Ihlamur, "Evaluation of early stage investments using nlp," *Vela Partners*, 2021.
- [27] M. Mitchell, *An Introduction to Genetic Algorithms*. The MIT Press, 1998, ISBN: 9780262280013. DOI: <https://doi.org/10.7551/mitpress/3927.001.0001>.
- [28] J. Schaffer, D. Whitley, and L. Eshelman, "Combinations of genetic algorithms and neural networks: A survey of the state of the art," in *[Proceedings] COGANN-92: International Workshop on Combinations of Genetic Algorithms and Neural Networks*, 1992, pp. 1–37. DOI: [10.1109/COGANN.1992.273950](https://doi.org/10.1109/COGANN.1992.273950).
- [29] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evol. Comput.*, vol. 10, no. 2, pp. 99–127, 2002, ISSN: 1063-6560. DOI: [10.1162/106365602320169811](https://doi.org/10.1162/106365602320169811).