Maverick: guiding venture capital investment through Deep Learning & NLP

Luka Secilmis & Yigit Ihlamur

August 17, 2022

Abstract

Maverick (MAV) is an AI-enabled algorithm to guide Venture Capital investment by leveraging BERT - the state-of-the-art deep learning model for NLP. Its ultimate goal is to predict the success of early-stage start-ups.

In Venture Capital (VC) there are two types of successful start-ups: those that replace existing incumbents (type 1), and those that create new markets (type 2). In order to predict the success of a start-up with respect to both types, Maverick consists of two models:

1. **MAV-Moneyball** predicts success of early stage start-ups of type 1.
2. **MAV-Midas** predicts whether a start-up fits current investment trends made by the most successful brand and long-tail investors, thereby taking into account new emerging markets that do not necessarily already have established successful start-ups leading them - ie. start-ups of type 2.

Maverick is developed through a transfer learning approach, by fine-tuning a pre-trained BERT model for type 1 and type 2 classification. In this paper we present Maverick, its development, and its performance. Notably, both MAV-Moneyball and MAV-Midas achieve a true positive ratio greater than 70%, which in the context of VC investment is one of the most important evaluation criteria as it is the percentage of successful companies that Maverick predicts to be successful.

1 Motivation

Vela Partners is a San Francisco-based investment firm, pioneering Venture technology (Ventech) - an innovative way of investing in startups using machine learning and data analytics. Vela is an AI-enabled data business, collecting insights about markets, companies, and other relevant components of the ecosystem. Then it leverages these insights to guide their venture capital investment decisions. In the world of investing textual data is abundant in the form of company descriptions, news, reports, contracts etc. To process and analyse this big data the use of NLP has become of utmost importance. NLP (Natural Language Processing) is a branch of artificial intelligence that uses machine learning to process, generate and understand natural language and its contextual nuances much in the same way humans do. In this project we will use NLP to gain insights from unstructured textual descriptions of companies by generating an understanding of the data using the state-of-the-art deep learning NLP model called BERT, and then constructing an additional layer on top of its neural network architecture for classification of type 1 and type 2 through transfer learning.

2 Maverick

Maverick (MAV) is an AI-enabled algorithm to guide venture capital investment by leveraging BERT - the state-of-the-art deep learning model for NLP. Its ultimate goal is to predict the success of early-stage start-ups, of which there are two types: (type 1) those that replace existing incumbents, (type 2) those that create new markets. In order to predict a start-ups success with respect to both types, Maverick consists of two models:

1. **MAV-Moneyball** predicts success of early stage start-ups of type 1.
2. **MAV-Midas** predicts whether a start-up fits current investment trends made by the most successful brand and long-tail investors, thereby taking into account new emerging markets that do not necessarily already have established successful start-ups leading them - ie. startu-ps of type 2.

Both MAV sub-models are developed through transfer learning by fine-tuning two pre-trained BERT models for type 1 and type 2 classification. In the following sections we will present Maverick's development and workflow.

2.1 Pre-trained BERT model

The first step of Maverick's workflow is to generate an understanding of the textual data fed into the algorithm. For this we use the deep learning NLP model BERT (Bidirectional Encoder Representations from Transformers) developed by Google. Specifically we will use the pre-trained BERT base (cased) model, one of the most revolutionary breakthroughs in the field of NLP, introduced by Devlin and colleagues from Google in their 2018 paper[1]. The model was trained in an unsupervised fashion for Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) on the entirety of Wikipedia and the Book Corpus dataset which contains around eleven thousand books. It is bidirectional and context-based allowing us to gain a deep understanding of language and its context. This model can then be fine-tuned, which is a powerful technique allowing us to leverage the pre-trained general language model, and specialize it in the task of binary classification.

2.2 Transfer Learning

The two datasets that are used to fine-tune Maverick for classification are the Moneyball and Midas dataset. The Moneyball dataset is composed of around 100 thousand company descriptions of which 11 thousand are successful (the start-up raised more than 50 million USD) and the rest unsuccessful (the start-up raised less than one million USD). This dataset is best suited for type 1 classification as already established successful startups will inherently not provide us information on newly emerging markets that we strive to predict with MAV-Midas for type 2 classification.

The Midas dataset is the result of Vela Partners' Midas Touch project which built a relationship graph between brand investors and long-tail investors used to decide if a long-tail investor is as reputable as a brand investor. (Long-tail investors are the thousands of small investors who are not as well known as top brand investors but their investment performance is similar). The dataset is made up of around 65 thousand company descriptions (we will call them ideas) of which around 3600 are ideas invested in by successful investors since 2018, and the rest are ideas/start-ups that have received no funding since their founding in 2018.

Due to the clear imbalance in successful and unsuccessful labels of the data two methods are employed: first descriptions of the unsuccessful label from both datasets are sampled to keep a two to one ratio of unsuccessful:successful companies. This is done to avoid the model from overfitting on the unsuccessful labels. Second, when splitting both datasets into training and test sets we use stratification to ensure that the 2:1 ratio of unsuccessful:successful labels is kept in both sets.

Finally, with help of the Transformers library[2], we employ transfer learning and fine-tune the pre-trained model for binary classification by adding an additional layer on top of the BERT-base-cased neural network architecture. This is done once with the Moneyball dataset and once with the Midas dataset creating two seperate models - MAV-Moneyball and MAV-Midas. Both models are trained for three epochs with ADAM - the same optimizer BERT was originally trained with.

3 Results & Success Criterion

To analyse the performance of both MAV-Moneyball and MAV-Midas we first investigate their classification accuracy on the test set. Then we compute the model's true positive ratio (TP). TP is one of the most important success criteria for Venture Capital as it is the percentage of successful companies that MAV indeed predicts to be successful.

- **MAV-Moneyball**
  - **Classification Accuracy:** 83%
  - **True Positive:** 77%
- **MAV-Midas**
  - **Classification Accuracy:** 66%
  - **True Positive:** 72%

4 Conclusion & Future Directions

This paper presented the development of Maverick - an AI-enabled and data driven algorithm, using NLP and deep learning, to guide venture capital investment. Through the use of transfer learning, Maverick is constructed of two models - MAV-Moneyball and MAV-Midas - that are engineered by respectively fine-tuning the pre-trained BERT-base cased model for type 1 (replacing existing incumbents) and type 2 (creating new markets) binary classification with two different datasets. The classification accuracy and true positive ratio, which we designate to be our ultimate success criterion in the context of VC investment, are computed. The results show to be very promising as both MAV-Moneyball and MAV-Midas achieve a TP ratio greater than 70%. As a result, Vela Partners will integrate this model in its production. Furthermore, the authors of this paper will work to launch Maverick to the public.

There are many future directions this project can take, such as increasing the data for type 2 classification. One approach for this could be creating a continuous data acquisition pipeline by which innovative ideas/concepts are scraped from sites such as TechCrunch for instance. This pipeline could also involve a new NLP algorithm which helps perform a form of sentiment analysis - labelling articles that are discussing innovation, new breakthroughs and potential market disruptions. In addition to the Midas Touch dataset this could increase the robustness of MAV-Midas or even render a new Maverick sub-model.

Continuity is a crucial concept to Maverick as markets, firms and technology are constantly evolving and changing. Thus, MAV must involve a form of continuous learning - which is the process of an AI algorithm learning on an on-going basis. This can for the moment be done by simply appending new data to the currently used Moneyball and Midas Touch datasets and fine-tuning from scratch again periodically. Currently this is a realistic solution as fine-tuning both models presented in this paper using one GPU is a relatively inexpensive and efficient operation in the order of 1-2 hours. As the datasets grow parallel-processing can be further used to reduce complexity of fine-tuning even more. However, researching different paradigms of continuous learning that allow us to avoid re-training completely could be useful in the long run.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[2] H. Face, "Transformers library." https://huggingface.co/docs/transformers/index.