AIR-mazing Predictions: Ranking of Retrieval Augmented Stock Market Prediction for Business Ideas

Kai Kainbacher, Maximilian Legat, Jonathan Maier, Michael Sickl
Group 09

January 2025

Advanced Information Retrieval
Institute of Human-Centred Computing

Project: https://github.com/jonnyCap/AIR-Project

1 Introduction

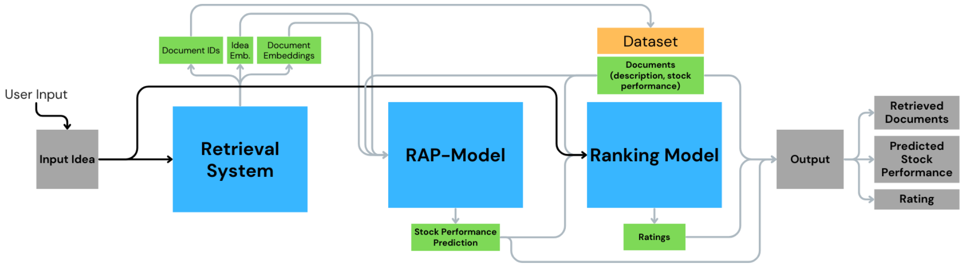
Trying to predict stock performance is very tempting due to its high rewards and has therefore been attempted many times. However, existing approaches typically focus on forecasting the performance of established companies, whereas our approach aims to provide insights into the potential of new business ideas. Motivated by the desire to build a system that provides initial feedback for new ideas we developed a system pipeline to address our core research question: "***Can textual descriptions of a new business idea be used to predict its stock performance and rank it alongside existing market competitors?***", we adopted a system pipeline approach that chains together three subsystems: the Retrieval System, the Retrieval-Augmented-Prediction Model (RAP-Model), and a Ranking Model. This report presents the results and evaluation of the entire system as well as its individual subsystems. Finally, a summarized conclusion is provided, answering our research question and a section on future work included at the end.

2 Related Work

While our system takes a novel approach by focusing on predictions for new business ideas rather than established companies, its architectural design remains closely aligned with traditional stock prediction systems. The architecture of our system was influenced by a study conducted by Pardeshi et al., 2023, where the authors proposed an LSTM model enhanced with a Sequential Self-Attention Mechanism (LSTM-SSAM), demonstrating significant improvements in prediction accuracy across multiple datasets. Inspired by this, we adopted a similar approach and built our architecture around LSTM and Self-Attention layers. Another key influence on this project was the growing recognition of retrieval-augmented techniques for enhancing forecasting systems. For example, Li et al., 2024 underscores the limitations of traditional machine learning methods in interpretability and reasoning. Similarly, Xiao et al., n.d. illustrates how retrieval-augmented approaches can address the complexities of time-series forecasting. While these studies primarily explore the use of LLMs with Retrieval-Augmented Generation (RAG), we adapted retrieval augmentation to an LSTM-based system.

3 Experiments and Results

To enable users to interact effectively with our system, we developed three subsystems that work together to deliver comprehensive information. Since each subsystem is interdependent, we designed a simple pipeline to seamlessly integrate the retrieval system, prediction-, and ranking-model.

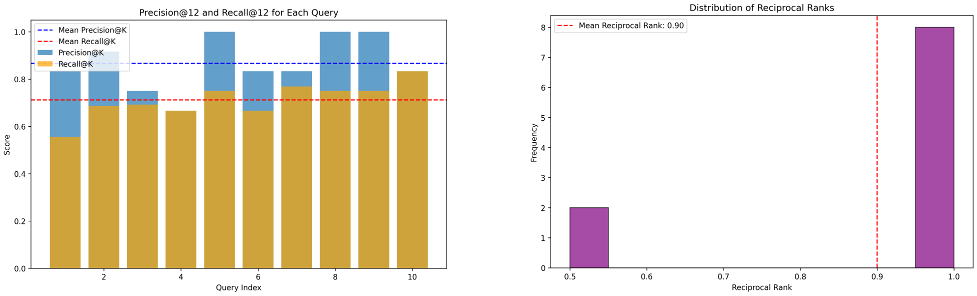


To enhance usability, we developed a GUI using C# Windows Forms, integrated with Python scripts for core tasks like retrieval, stock prediction, and ranking, that allows for direct interaction with our models through the C# interface. More details are available in our repository's README.md.

Considering that our system is composed of three subsystems, each subsystem will first be evaluated individually, followed by an assessment of the entire system as a whole. The notebooks for training, evaluation, as well as our datasets, can be found in our repository.

3.1 Retrieval System

Starting with the Retrieval System, it employs a **BERT Encoder** to transform input text into embeddings following critical preprocessing steps. Using **cosine similarity**, the system identifies the most similar businesses from our dataset. To evaluate its performance, we conducted a manual analysis by generating ideas, manually identifying similar companies, and comparing these with the system's output. We calculated key metrics, including **Precision**, **Recall**, **Mean Reciprocal Rank**, **Mean Average Precision**, and **Normalized Discounted Cumulative Gain**. Our evaluation showed that the system performs exceptionally well, demonstrating its ability to retrieve most relevant documents.



(a) Precision & Recall

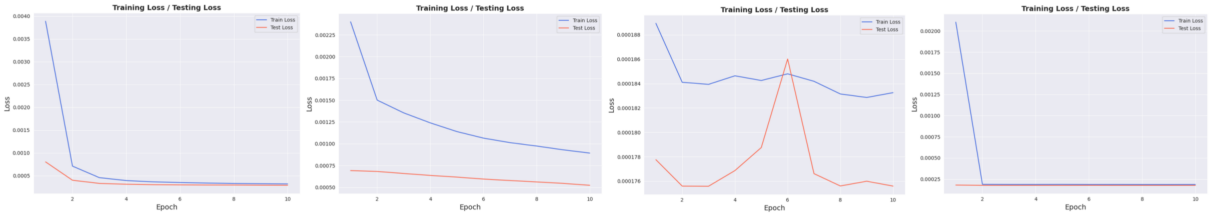
(b) Mean reciprocal rank

Figure 1: Retrieval System: Insight into metrics

In these images you can see that average precision and recall is very high and that the system found for almost every query the most important document.

3.2 Retrieval Augmented Prediction Model

The RAP-Model leverages retrieved documents as inputs to enhance predictions of an idea's potential stock performance. Given the challenges of predicting stock trends, we developed multiple architectures incorporating attention, LSTM, and autoregressive predictions, each configurable for specific needs. The most promising results came from the **Retrieval Augmented Prediction Model** without attention optimizations, leading to further testing with tailored configurations. Two models utilized **auxiliary inputs** with either **Custom Loss** or **SmoothL1Loss**. The Custom Loss combined Temporal Loss, MSE Loss, a penalty for negative values, and Huber Loss for robustness against outliers. Two additional models excluded auxiliary inputs with the same loss variations. (Link to training notebook).



(a) No Auxiliary & Custom Loss

(b) Auxiliary & Custom Loss (10 Epochs)

(c) Auxiliary & Custom Loss (20 Epochs)

(d) Auxiliary & SmoothL1Loss

Figure 2: RAP-Model: Loss during training

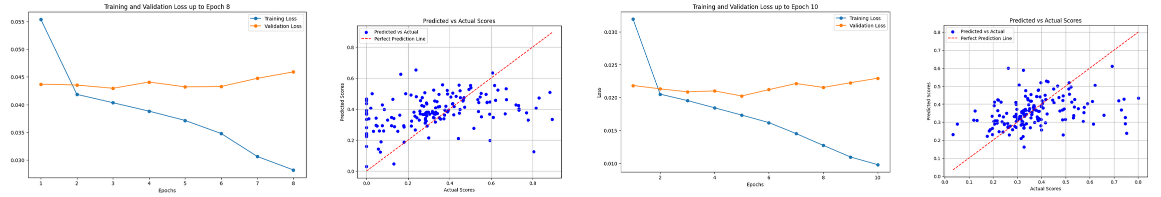
Due to technical issues, results for the model without auxiliary input functionality using SmoothL1Loss are excluded. In Image (d), the model converges quickly within two epochs but shows no further improvement, while in Image (a), the loss curve also stabilizes after a few epochs. The most notable performance comes from the model using auxiliary input functionality with custom loss. As seen in Image (b), this model required more training to converge, why training was extended by 10 epochs (Image (c)), during which the loss initially increased but later resumed its downward trend.

3.3 Ranking-Model

The Ranking-Model purpose is to assign a score to new ideas and their predicted stock value, which is then ranked among similar competitors. For properly training this model multiple metrics were created, that should represent an idea's potential. Eventually we came up with a metric that does this very well was assessed as sufficient in a qualitive sample evaluation. It can be computed as follows:

$$\text{Score} = \left(0.4 \times \frac{S_{\text{norm}} \times M_{\text{norm}}}{0.8 + |V_{\text{norm}} - 0.5|} + 0.4 \times D_{\text{norm}} + 0.4 \times \sqrt{M_{\text{norm}} \times V_{\text{norm}}}\right)$$

Unfortunately, a pattern emerged where seemingly good stock predictions received low ratings, as the model apparently linked strong closing prices with poor performance indicators, resulting in a lack of transparency. To address this, we optimized the formula to better align scores with stock performance, which can be found in our repository, calculated as score-Nr.:5. While this adjustment slightly reduced score accuracy, it made the results more user-friendly. Regardless of the formula used, the training process effectively minimized loss on the training data, while the evaluation loss plateaued, suggesting a lack of correlation between business descriptions, recent stock performance, and the calculated score:



(a) Training & Validation Loss (Score 2)

(b) Predicted vs. Actual Scores (Score 2)

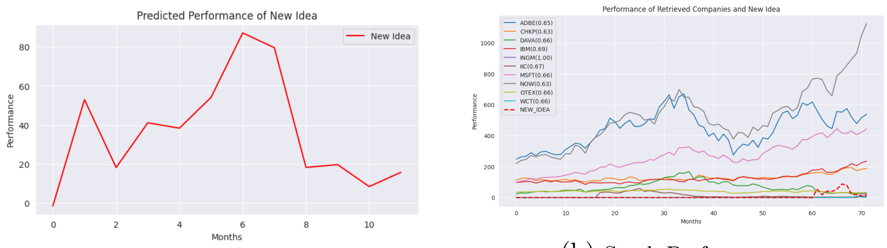
(c) Training & Validation Loss (Score 5)

(d) Predicted vs. Actual Scores (Score 5)

Figure 3: Ranking Model: Loss during training

3.4 Overall Results

In addition to quantitative analyses on the evaluation data, we conducted a qualitative analysis of several promising companies to assess the realism of predictions, while recognizing that certain real-life events may not be accounted for. Details of the evaluation are available in our Evaluation Folder. A notable limitation of the models is their inability to handle nonsensical input. Additionally, the RAP-Model exhibits bias, evident when comparing its performance across all models, as detailed in this comparison notebook. Predictions consistently range between \$20 and \$100, likely influenced by the mean value of all predictions. Despite its limitations, the model performed well in some cases. For instance, it predicted a stock value of \$20 for Ingram Micro Holding Corporation, a company not in our dataset, with an actual value on January 11, 2025, ranging from \$19.11 to \$19.98. This low value aligns with its rank of seventh out of eleven in our ranking model.



(a) No Auxiliary & SmoothL1Loss Prediction

(b) Stock Performances

Rank	Ticker	Ranking
6.	CHKP	0.4774
7.	NEW IDEA	0.4562
8.	ADBE	0.3890

Table: Model Ranking

Figure 4: Prediction and Ranking with Images and Table

As one of many positive examples, this showcases that our pipeline in fact is capable of providing accurate and realistic information about a new business idea's potential.

4 Conclusion

In conclusion our System works well and does provide valuable insights into a new business idea's potential, as it successfully retrieves competitors, predicts stock performance within realistic boundaries and provides a ranking system that can detect success factors at a reasonable rate. However, there are still multiple flaws and room for extending and improving each subsystem, especially our RAP- and Ranking Model. In the following challenges, key takeaways and future work will be further discussed.

4.1 Discussion & Challenges

The dataset used for training the RAP-Model faces several challenges: its small size limits diversity, and business descriptions often include irrelevant information like locations or management strategies, leading to unintended contextual biases. Additionally, the dataset lacks an optimized balance between active and failed companies, which likely skews predictions, causing the model to rarely assign low scores and potentially biasing its perception of failure. These dataset limitations directly impact the RAP-Model's performance, compounding its own inherent challenges. The model is highly sensitive to small changes due to its autoregressive nature, making its accuracy heavily reliant on the choice and design of the loss function. Furthermore, it often overfits to patterns from recent months, predicting past trends rather than accurately forecasting future outcomes. Coupled with the dependence on an up-to-date dataset, these issues highlight the critical importance of regular updates to ensure accurate predictions. Although impacted by the dataset limitations, the Ranking Model faced additional challenges with uncorrelated inputs and outputs, as well as a lack of transparency. However, the application of appropriate methodologies significantly improved the interpretability of its scores. Despite using a less optimized metric, the model ultimately delivered reasonable and practical results.

4.2 Key Takeaways

A very important takeaway is that that information retrieval can indeed enhance stock predictions, even for completely new ideas that have not yet entered the market, as indicated by our results. Another significant takeaway is the critical role the dataset plays in a model's performance. Additionally, the findings emphasize the importance of transparency in prediction models. While enhancing interpretability may sometimes reduce accuracy, the trade-off is worthwhile when it results in outputs that are more reliable for users. This highlights the importance of designing models that balance predictive power with clarity and usability. Furthermore, the research underscores the impact of choosing the right architectural configurations and loss functions. The experiments with multiple versions of the RAP-Model demonstrate that exploring diverse configurations can lead to models that better adapt to various input conditions. Lastly, and most importantly, our core research question is answered: **Yes, a business description can be used to predict and rank a business's potential on the stock market.** However, this conclusion holds only if the model is used with realistic inputs.

4.3 Future Work

The system's potential is substantial, with opportunities for optimization and enhancement. Integrating large language models (LLMs) or improving the attention-optimized RAP-Model shows promise for enhancing its capabilities. To address the insufficient correlation between inputs (text embeddings, stock performance) and outputs (idea scores) in the Ranking Model, strategies like developing better evaluation metrics or adopting alternative modeling approaches could be explored. For example, a parallel evaluation model that scores all documents simultaneously could improve coherence and consistency, resulting in more reliable and actionable rankings.

References

Li, X., Li, Z., Shi, C., Xu, Y., Du, Q., Tan, M., Huang, J., & Lin, W. (2024). Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*.

Pardeshi, K., Gill, S. S., & Abdelmoniem, A. M. (2023). Stock market price prediction: A hybrid lstm and sequential self-attention based approach. *arXiv preprint arXiv:2308.04419*.

Xiao, M., Jiang, Z., Chen, Z., Li, D., Chen, S., Ananiadou, S., Huang, J., Peng, M., & Xie, Q. (n.d.). Timerag: It's time for retrieval-augmented generation in time-series forecasting.

List of Figures