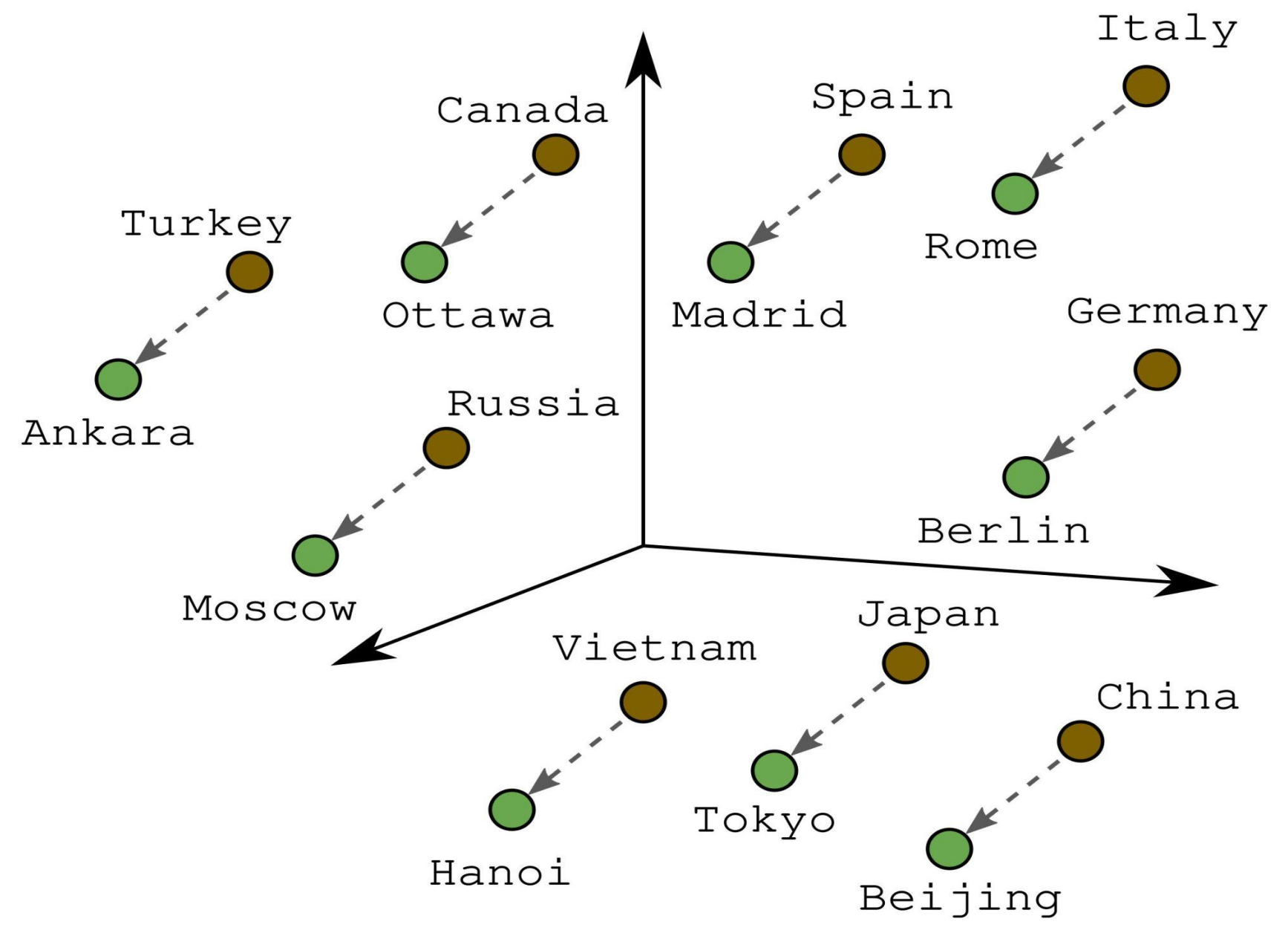
Stock2Vec:Optimizing Predictions of Company Characteristics with Stock Embedding

Zhelu Mai[1], Diya Deepak[2], Sugam Kafle[3], Madhav Thamaran[4], Ting Xiao[5], Mark V. Albert[5]
Department of Computer Science and Engineering[2], University of North Texas
G. Brint Ryan College of Business[1,3], University of North Texas | Texas Academy of Mathematics and Science[4]

Abstract

Price changes in stocks play a significant role in many ways. Not only can it capture the similarity between companies, but it can also dive deeper into hidden characteristics like weather conditions. For example, hurricanes in a particular region can also impact the stock price changes of companies in that region. Thus, it's valuable to create an Embedding of company stocks, Stock2Vec, which can be easily added on to and optimize any Prediction Model that applies to companies with associated stock prices. In our work, based on the framework of the previous Stock2Vec paper, we built a fine-tuned Word2Vec and FastText model. Using three different regression models, we predicted a new target variable, the Market Capital of companies, with three different Embedding Models. The experiment results demonstrate that our models achieved at least a 10% improvement in performance.

The Value of Embeddings



Country-Capital

Figure 1: Graph Representation of vector embeddings. Similar vectors are grouped together.

Embeddings: Vector representations of real world phenomena

Types of Embeddings (Everything can be 2Vec!):

- Place2Vec
- Doc2Vec
- Word2Vec
- Stock2Vec

Results & Findings

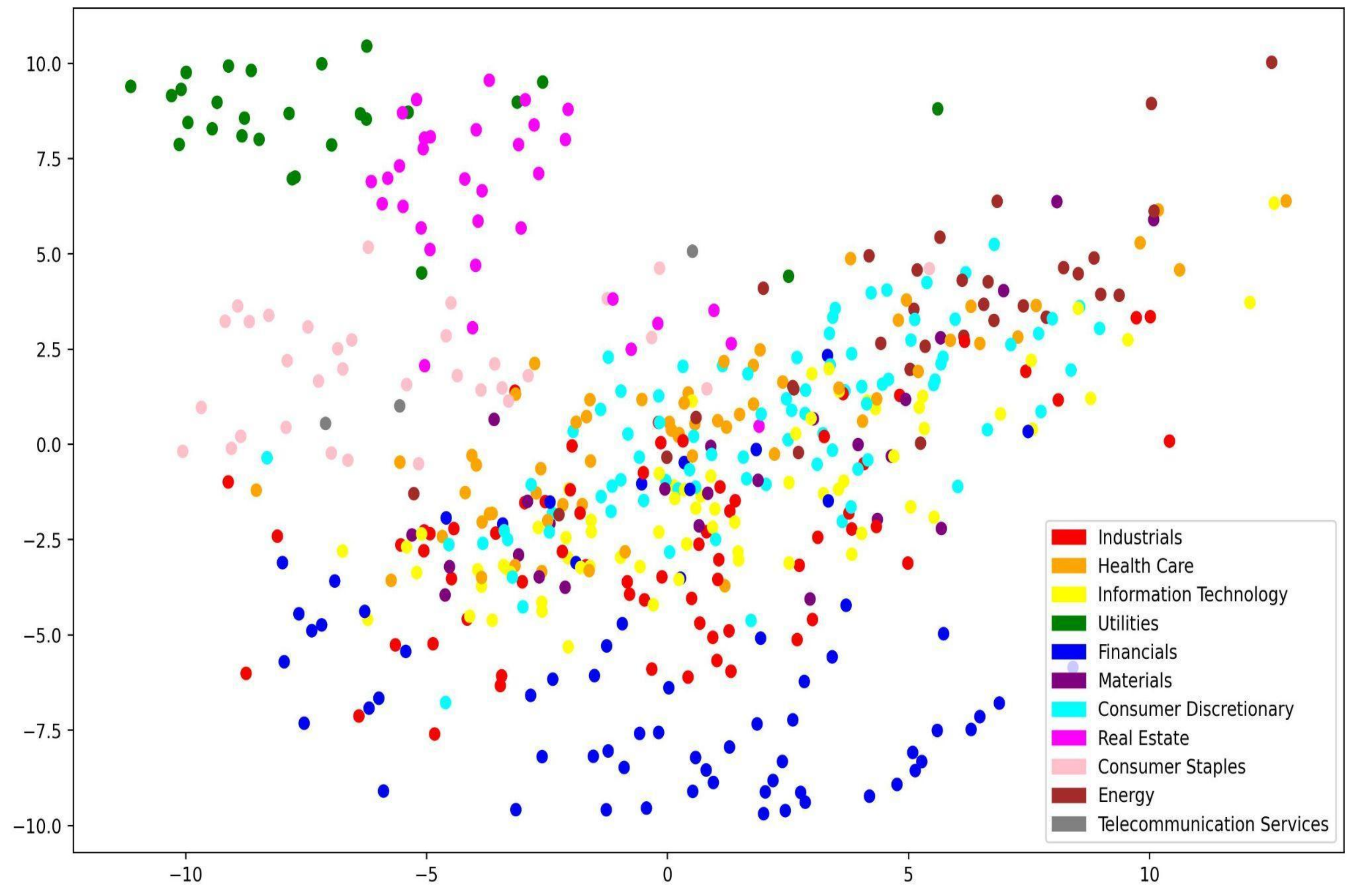


Figure 2: PCA plot of companies by color coded by corresponding sectors

- Utilized the fine-tuned Word2Vec model
- Clusters were less widely distributed and overlapped than existing model
- Better clustered data points than the existing model

In the same prediction task (**ESG ratings)**:

- the **fine-tuned Word2Vec** model performed best in the task
- **Random Forest Regression** got the highest R squared score in the task

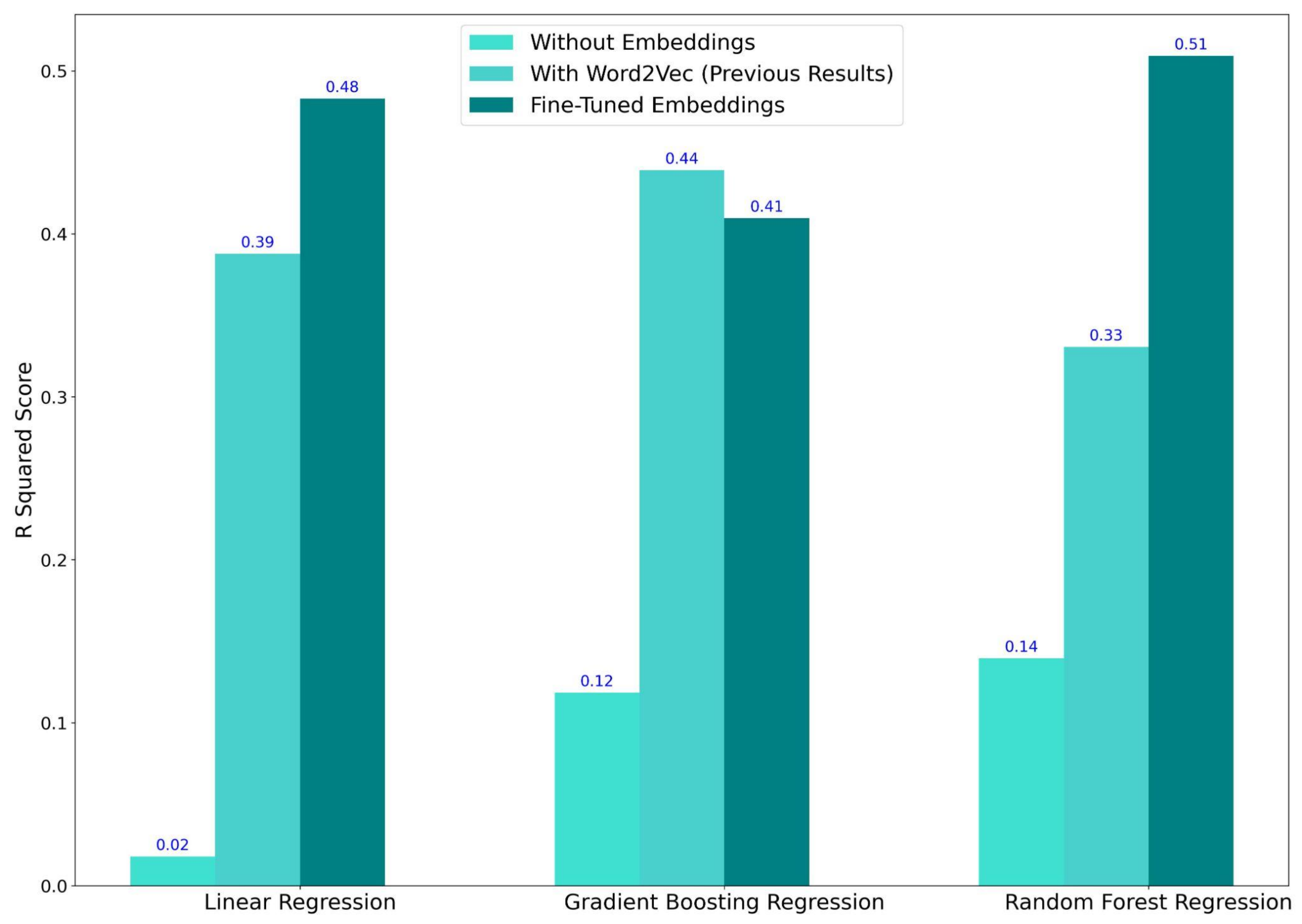
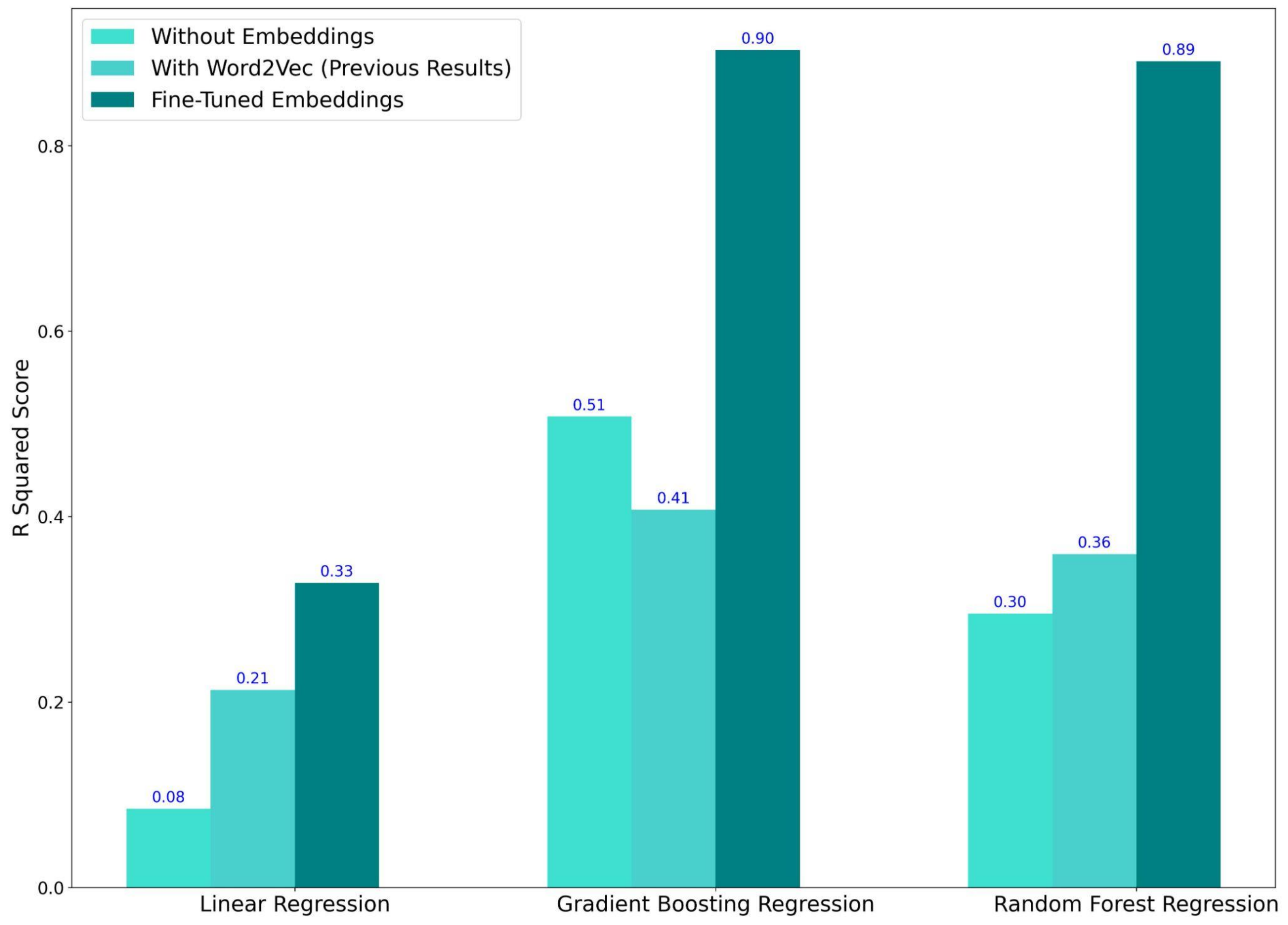


Figure 3: Performance Comparison of Predicting ESG Ratings

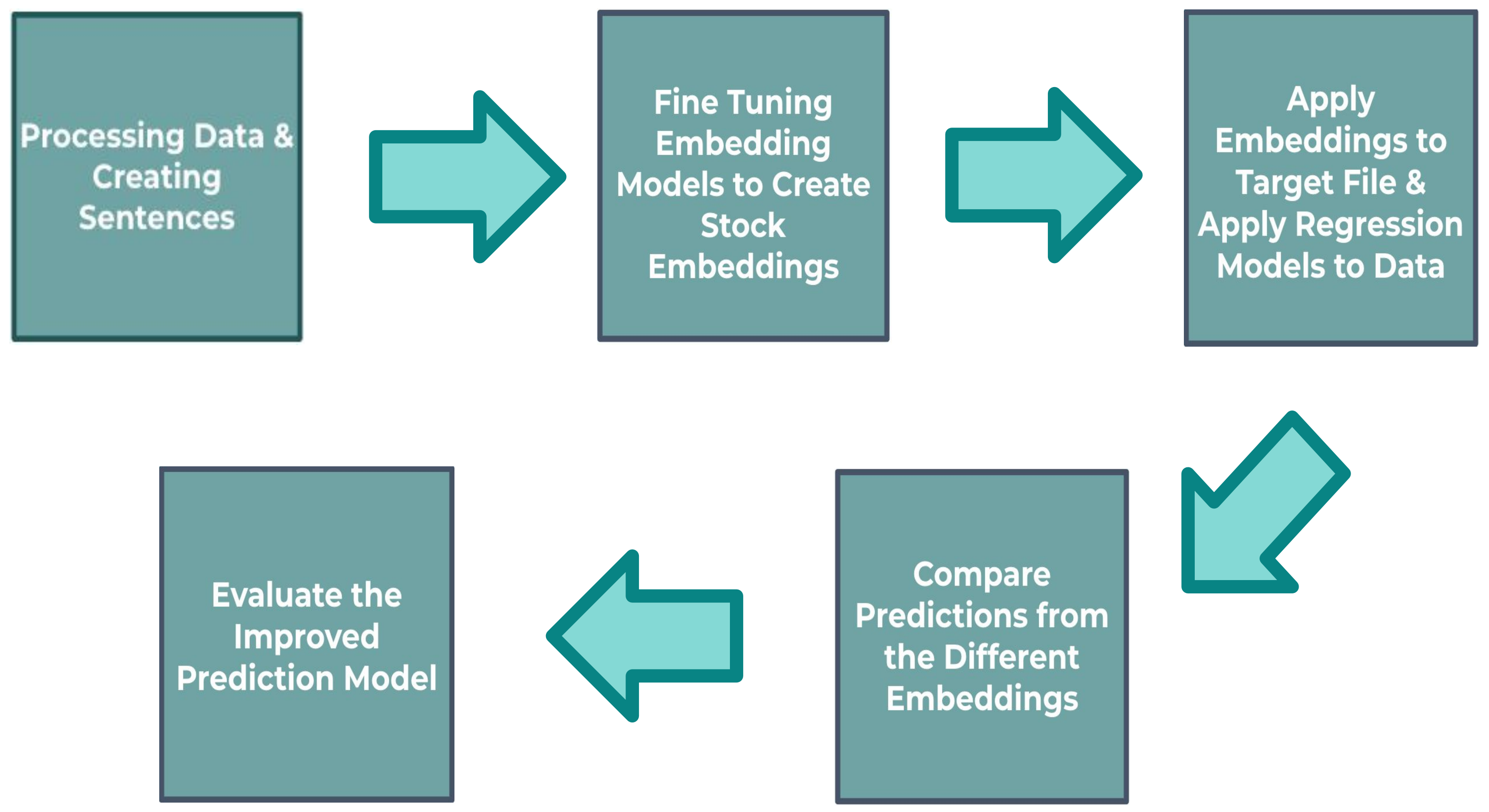


In predicting the **Market Capital** of SP 500 Companies:

- the **fine-tuned Word2Vec** model did a better job in new task
- **Linear Regression** reached the worst result and others regressors conducted barely the same

Figure 4: Performance Comparison of Predicting the Market Capital of SP 500 Companies

Methodology



Processing Data & Creating Sentences → Fine Tuning Embedding Models to Create Stock Embeddings → Apply Embeddings to Target File & Apply Regression Models to Data → Compare Predictions from the Different Embeddings → Evaluate the Improved Prediction Model

How we modified the Word2Vec and FastText model:

- Set the word window to 50
- Applied negative sampling
- Raised the dimensions from 4 to 10 (FastText) and 12 for (Word2Vec)

Conclusion & Future Work

To Summarize:

- Compared to the PCA, better clustered data points increased performance
- Random Forest Regression outperformed Linear Regression in predicting ESG and Market Capital,
- Fine-tuned Word2Vec and FastText embeddings work better than the previous model in predicting ESG and Market Capital, meaning they can adapt general taks.

Future Work:

- New input and new embedding models are needed because the Word2Vec and FastText models cannot capture the ordering of the sentence
- More data is needed because only five years of data may not capture the transformation of companies and sectors in the long term
- Fine-tuning the model using a dense layer is welcome

Acknowledgements

We would like to acknowledge authors of the previous paper on which this work is based: **Yuvraj Baweja, Phillip Nelson, George Mihaila, JagdishKumar Katariya**

Yi, Ziruo, et al. "Stock2Vec: An Embedding to Improve Predictive Models for Companies." arXiv preprint arXiv:2201.11290 (2022).