Scientific Research Publishing

Theoretical Economics Letters, 2025, 15(1), 187-198
https://www.scirp.org/journal/tel
ISSN Online: 2162-2086
ISSN Print: 2162-2078

Semantic Diversification in Equity Portfolios

Crina Pungulescu

Department of Economics, John Cabot University, Rome, Italy
Email: cpungulescu@johncabot.edu

How to cite this paper: Pungulescu, C. (2025). Semantic Diversification in Equity Portfolios. *Theoretical Economics Letters, 15,* 187-198.
https://doi.org/10.4236/tel.2025.151011

Received: October 11, 2024
Accepted: February 7, 2025
Published: February 10, 2025

Abstract

In the race to harvest the power of Artificial Intelligence (AI) in virtually every field, researchers and practitioners are faced with an ever increasing supply of novel tools that have not undergone domain-specific tests. This paper informs the methodological choices of researchers in economics and finance by comparing the performance of three Natural Language Processing (NLP) methods at an important task, namely using text analysis for portfolio diversification. Portfolio management can benefit from analysing text data in the form of company descriptions, since the returns of companies with similar descriptions tend to be correlated and consequently, portfolios of dissimilar companies should have lower risk. In this paper, three NLP methods are used to construct so-called minimum semantic concentration portfolios, which are designed to leverage the semantic diversity of the business descriptions of constituent companies to reduce portfolio volatility. Two widely used large language models (BERT and GPT) and an alternative AI solution inspired by neuroscience, called semantic fingerprinting are put to the test of comparing meaningfully the business descriptions of the S&P 500 and respectively Europe 600 constituents in order to derive actionable investment insights. The results show that all three NLP methods are able to extract relevant information from company descriptions: the minimum semantic concentration portfolios have significantly lower volatility than portfolios constructed with randomly chosen weights. While no NLP method is able to claim absolute superiority over its peers, semantic fingerprinting appears the most consistent and robust performer, since BERT and GPT demonstrate not only their potential but also a caveat, as their performances are volatile even across very similar tasks.

Keywords

Text Analysis, Portfolio Performance, Natural Language Processing, BERT, GPT, Semantic Fingerprinting

1. Introduction

"*Data is the new oil*"[1]. Forbes provides us with a powerful image for data as a valuable business asset and the condition for success is equally aptly described: "*but one needs a powerful engine to extract, refine and harness it efficiently*". Unstructured data (in the form of text, images, audio and video) appears to be the fastest growing form of data in the last years[2] and powerful engines are already competing in extracting valuable information from large data pools. The ability to leverage previously unexplored information is due to important achievements in the field of artificial intelligence (AI) where Natural Language Processing (NLP) techniques make it possible for computers to analyse human language with high-level accuracy. Traditional NLP algorithms designed to extract meaning from vast quantities of text data have recently been augmented by large language models (LLMs), capable to generate coherent context and even engage in human-like conversations.

Large language models, such as Google's BERT and OpenAI's GPT obtain spectacular results and are already widely used. Their "magic" is an impressive feat of engineering and statistics, as their predictions involve hundreds of billions (in some cases, trillions) of parameters. An alternative, less computationally intensive AI solution that is able to "read text" relies, on the other hand, on a theoretical model of the brain. Cortical.io has developed its *Semantic Folding Theory of Language Data Representation* (De Sousa Webber, 2016) based on Numenta's *Thousand Brains Theory of Intelligence*, which identifies the cortical column as the basic functional unit (tightly replicated across the cortical sheet) and states that all cortical columns, even in low-level sensory regions, are able to learn and act as thousands of brains working simultaneously (Hawkins et al., 2017). The result is that any concept is represented in the brain by the activation of a small subset of hundreds of thousands of available cortical columns to create map-like reference frames. For example, the concept of a face is defined in the brain as the combination of its elements (nose, eyes etc.) and their relative spatial positions. In this way, the brain has developed an efficient algorithm to model sparse distributed representations of every meaning (Hawkins, 2021). Semantic folding applies this model of the brain to language representation, by efficiently converting any term or text into a symbolic numerical format, a sparse binary vector, called the semantic fingerprint. In the training phase, the semantic space consisting of 16,384 contexts is created using word associations "learned" from a reference corpus (Wikipedia in this case). Each context is a bag-of-words that have been frequently co-occurring within the reference corpus. Once the semantic space has been constructed by defining the 16,384 contexts, any given term or text can be placed within this space based on the frequency of its co-occurrences with any of the pre-defined contexts. Thus, the semantic fingerprint of any input text is a binary vector of length equal

[1]https://www.forbes.com/sites/forbestechcouncil/2023/01/11/five-data-analytics-trends-on-tap-for-2023/?sh=755cbc756cfd.

[2]https://www.datanami.com/wp-content/uploads/2019/01/Zorroa_data_growth.png.

to 16,384 bits (the number of contexts in the semantic space), with a subset of 984 contexts that are most relevant (i.e. frequently associated with the input text within the reference corpus) highlighted by receiving the value 1[3].

As both text data and methods of extracting it efficiently are becoming increasingly available, what is missing in economics and finance are field-specific validation tests (see Ash and Hansen, 2023, for an in-depth treatment of the current limitations of text algorithms in economics). This paper provides one such test by comparing the ability of three NLP methods—two popular large language models as well as semantic fingerprinting, first applied to finance by Ibriyamova et al. (2017)—in the context of portfolio diversification, a fundamental principle in finance. The theoretical underpinnings of using text analysis to achieve lower correlation and therefore lower overall risk in a portfolio of financial assets are provided by Ibriyamova et al. (2019) who show that obtaining portfolio weights that minimise (subject to the budget constraint) the semantic concentration (or semantic similarity of business descriptions) of the constituent assets of a portfolio, is mathematically identical to the process of identifying the minimum variance portfolio. Related papers follow the methodology of Ibriyamova et al. (2017, 2019) and compare the performance of current NLP methods in the context of predicting return correlations for U.S. (Pungulescu and Stolin, 2023) and respectively European stocks (Pungulescu, 2024). This paper compares the same three NLP methods in terms of the diversification benefits they bring (i.e. lower portfolio risk) using a dataset consisting of the largest U.S. and European companies and spanning more than two decades to construct semantically diversified (i.e. minimum semantic concentration) portfolios of 25 and respectively 50 companies. The results show that all three NLP methods lead to semantically diversified portfolios whose returns are significantly less volatile than those of portfolios with randomly chosen weights. However, none of the methods stands out as a clear frontrunner in terms of performance. Semantic fingerprinting (SF) emerges as a dependable and cost-effective choice, delivering consistently high performance. In contrast, BERT and GPT, while offering viable solutions, exhibit greater variability in performance, even in tasks that are extremely similar. This emphasises the importance of making an informed choice when considering particular variants of large language models, which might involve a preliminary research stage of extensive comparisons and potentially fine-tuning the models. Given the substantial costs (financial and environmental[4]) that model training, fine-tuning and comparison entail and the fact that higher costs do not always translate into significantly higher performance[5], task-driven efficiency becomes an important decision criterion. A "small" model might perform a specific task with suitable accuracy and for a

[3]For a more detailed description of semantic fingerprinting, see Pungulescu (2022a).

[4]Dodge et al. (2022) estimate that training a 6 billion parameter Transformer model would use 103.5 MWh of electricity, that is 2800 times more than training a BERT Small model.

[5]Chen et al. (2023) compare their own solution, FrugalGPT (a combination of various large models) to OpenAI's GPT-4 and report a cost reduction of 80%, while improving accuracy by 1.5%.

fraction of the cost.

The remainder of the paper is organised as follows. Section 2 presents the data, section 3 lists the NLP methods employed, while section 4 presents the results and section 5 concludes.

2. Data

Daily price data for the S&P 500 and respectively Europe 600 constituent companies is retrieved from Bloomberg for the period 3.01.2000-9.03.2022. The 1100 companies in the dataset represent approximately 80% of the total market capitalisation of U.S. public companies and 90% of the European stock market (not limited to the Eurozone) and therefore enable a large scale test of the benefits of semantic diversification for portfolio creation.

Since a unique provider for all needed business descriptions has not been available, several sources have been used for data collection, retaining the longest text for the companies that were covered by more than one database. For the S&P 500 constituents, the business descriptions are obtained from Refinitiv Eikon (145 companies), Reuters (215 companies) or Forbes (140 companies). The business descriptions are between 83 and 421 words long, with an average of 145 words per description (and standard deviation of 30 words). For the Europe 600 constituent companies, the sources are Refinitiv Eikon (138 companies), GlobalData (61 companies), and Yahoo Finance (400 companies). The descriptions of the European companies are on average longer than those of the U.S. companies, at 166 words per description (standard deviation of 54 words) and within a range of 54 to 601 words.

3. Natural Language Processing (NLP) Methods

Following previous work by Pungulescu and Stolin (2023) and Pungulescu (2024), the business descriptions (in their original format) of the S&P 500 and Europe 600 constituent companies are quantified (i.e. converted into vectors of varying lengths) using the following NLP methods:

- Cortical.io's semantic fingerprinting (SF) technology using the "*Retina*" engine[6], which outputs a sparse binary vector where 984 out of a total of 16,384 positions are activated (with a value of 1).
- Google AI Language's Bidirectional Encoder Representations from Transformers (BERT) introduced by Devlin et al. (2019) in three variants: the "*bert-base-nlimean-tokens*" model, which converts each business description into a dense vector of embeddings of 768 positions, the "*all-MiniLM-L6-v2*" model (384 positions) and finally the "*all-roberta-large-v1*" (1024 positions).
- OpenAI's text similarity "*davinci-001*" engine[7] using the GPT-3 tokeniser (GPT), which converts each business description into a dense vector of 12,288 positions.

[6]https://www.cortical.io/science/sparse-distributed-representations/?highlight=retina.
[7]https://openai.com/blog/introducing-text-and-code-embeddings.

4. Methodology and Empirical Analysis

Cosine Similarity

Following the established methodology for measuring document similarity (see Ash and Hansen, 2023), company similarity is measured as the cosine similarity of the numerical vectors to which the companies' descriptions have been converted (using in turn each of the NLP methods introduced in Section 3):

$$\cos\left(v_A, v_B\right) = \frac{v_A^T v_B}{\|v_A\|\|v_B\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{1}$$

where the vectors $v_A = (a_1, a_2, \cdots, a_n)$ and $v_B = (b_1, b_2, \cdots, b_n)$ correspond to the vectorised business descriptions of company A and company B, respectively. Table 1 and Table 2 present descriptive statistics for the similarity measure for the S&P 500 and respectively the STOXX Europe 600 companies and illustrate the potential for divergence (and contest) across the three NLP methods. Pairwise correlations are positive for all the methods but moderate: for the S&P 500 companies in the 0.43 - 0.69 range and similarly, the range is 0.41 - 0.65 for the European companies.

The proven (Ibriyamova et al., 2017, 2019) and persistent (Pungulescu and Stolin, 2023; Pungulescu, 2024) link between pairwise return correlations and the similarity of the respective companies' business descriptions (with the returns of more similar companies being more correlated), can be leveraged by constructing minimum semantic concentration portfolios (see Ibriyamova et al., 2019, for the theoretical exposition). The minimum semantic concentration portfolio weights

Table 1. S&P 500: Cosine similarities.

	SF (16,384)	BERT (768)	BERT (348)	BERT (1024)	GPT (12,288)
			Descriptive Statistics		
Mean	0.36	0.58	0.23	0.41	0.75
Std.	0.08	0.10	0.10	0.10	0.03
Min.	0.08	0.20	−0.12	−0.05	0.56
Max	0.78	0.96	0.78	0.90	0.92
			Pairwise Correlations		
SF (16,384)	1.00	0.57	0.55	0.48	0.48
BERT (768)		1.00	0.62	0.69	0.48
BERT (348)			1.00	0.55	0.50
BERT (1024)				1.00	0.43
GPT (12,288)					1.00

The table reports descriptive statistics (mean, standard deviation, minimum and maximum) of the pairwise cosine similarities of the business description of the S&P 500 companies, as well as their correlations.

Table 2. STOXX Europe 600: Cosine similarities.

	SF (16,384)	BERT (768)	BERT (348)	BERT (1024)	GPT (12,288)
			Descriptive Statistics		
Mean	0.34	0.60	0.24	0.25	0.74
Std.	0.09	0.09	0.09	0.09	0.03
Min.	0.06	0.25	−0.09	−0.08	0.57
Max	0.78	0.96	0.82	0.85	0.93
			Pairwise Correlations		
SF (16,384)	1.00	0.52	0.45	0.41	0.48
BERT (768)		1.00	0.49	0.50	0.54
BERT (348)			1.00	0.65	0.50
BERT (1024)				1.00	0.54
GPT (12,288)					1.00

The table reports descriptive statistics (mean, standard deviation, minimum and maximum) of the pairwise cosine similarities of the business description of the STOXX Europe 600 companies, as well as their correlations.

are given by $\frac{\left(S^T S\right)^{-1} 1}{1^T \left(S^T S\right)^{-1} 1}$, where $S^T S$, the matrix of cosine similarities, takes on the role that the variance-covariance matrix, Σ, plays in finding the minimum variance portfolio weights. Pungulescu (2022b) compares the minimum variance and the minimum semantic concentration portfolios constructed from the Dow Jones Industrial Average constituent companies over 16 years and finds that the minimum semantic concentration portfolio outperforms the minimum variance portfolio 66% of the time.

In this paper, three NLP methods are placed in a "horse race" that attempts to find which of them results in a minimum semantic concentration portfolio with lower variance. Similar comparisons have been conducted in the related literature for different text analysis methods to show both that text analysis is apt to extract meaningful information from unstructured data as well as to rank the various methods. For instance, Ash and Hansen (2023) compare 10 different algorithms in terms of their ability to quantify similarities between companies using the Risk Factors section of the 10-K reports (required by the U.S. Securities and Exchange Commission) as input text, and Ibriyamova et al. (2017) task 4 measures of proximity for pairs of companies with predicting stock return correlations and find that the measure based on semantic fingerprinting performs best.

To test the ability of each of the three NLP methods to provide diversification benefits, for each day, a set of 25 and respectively 50 constituent companies of

S&P 500 and STOXX Europe 600 indices is randomly selected and minimum semantic concentration portfolios are constructed based on the cosine similarities of their business descriptions vectorised using each of the NLP methods, in turn[8]. These portfolios are held for the next 60, 120 and 240 days and the standard deviations of their returns are calculated over these holding periods. Time series of standard deviations of the returns of the portfolios' constructed on any given day are obtained in this fashion and the procedure is repeated 100 times. Table 3 and Table 4 compare the performance of 100 times series of standard deviations of returns of portfolios constructed using 25 constituents of the S&P 500 and respectively the STOXX Europe 600 indices. The comparison includes a portfolio with randomly chosen weights (RND) and five minimum semantic concentration portfolios based on our NLP methods: semantic fingerprinting (SF), three variants of BERT (resulting in vectors of differing lengths), and GPT. The results show how many times (out of 100), the time series of standard deviations of returns of portfolios with weights computed using a given method (SF, BERT, GPT, RND) and held for a particular number of days (60, 120, 240) has a lower mean (statistically significant at 10%) than the other methods. As expected, the portfolios using randomly chosen weights are without exception 'beaten' by all the minimum semantic concentration portfolios, illustrating the potential diversification benefits of using text analysis in the context of portfolio management. For portfolios of 25 constituent companies of the S&P 500 (see Table 3), using BERT 328 results in portfolios with a higher number of "wins" (counts with time series of standard deviations of portfolio returns that are significantly lower, on average, than for the portfolios constructed using the other methods). A close second is the NLP solution provided by GPT followed by semantic fingerprinting (SF). The two remaining BERT variants (with vectors of 1024 and respectively 768 positions) are in the last places. For portfolios of 25 constituent companies of the STOXX Europe 600 (see Table 4), BERT 1024 and semantic fingerprinting (SF) lead in a closely contested race (they are only one "win" apart), BERT 328 is a distant third and GPT comes in second to last. The difference between the performances of the BERT variants is suggestive of one of the potential pitfalls of using pre-trained models, as the best choice is by no means evident *a priori* and the wrong choice is clearly costly. The end-user is faced with the task of running extensive comparisons and/or further training and fine-tuning any chosen model.

For portfolios of 50 constituent companies of the S&P 500 (see Table 5), SF results in portfolios with the same number of "wins" as BERT 328 and they both outperfom GPT (albeit slightly), whereas for European companies, semantic fingerprinting (SF) wins clearly, followed at considerable distance by BERT 1024 (see Table 6). GPT manages to "beat" BERT 328 on this occasion and BERT 768 fares worst.

[8]The chosen portfolio sizes are in line with the recommendations of Alexeev and Tapon (2013) for developed markets.

Table 3. S&P 500: Minimum semantic concentration portfolios: 25 companies.

		RND	SF	BERT	BERT	BERT	GPT	**Wins**
	RW		(16,384)	(768)	(348)	(1024)	(12,288)	**Total**
RND	**60 days**	0	0	0	0	0	0	**0**
	120 days	0	0	0	0	0	0	**0**
	240 days	0	0	0	0	0	0	**0**
SF (16,384)	**60 days**	93	0	100	0	100	0	**293**
	120 days	92	0	100	0	100	0	**292**
	240 days	92	0	100	0	100	0	**292**
BERT (768)	**60 days**	93	0	0	0	0	0	**93**
	120 days	92	0	0	0	0	0	**92**
	240 days	92	0	0	0	0	0	**92**
BERT (348)	**60 days**	93	9	100	0	100	0	**302**
	120 days	92	30	100	0	100	0	**322**
	240 days	92	92	100	0	100	0	**384**
BERT (1024)	**60 days**	93	0	100	0	0	0	**193**
	120 days	92	0	100	0	0	0	**192**
	240 days	92	0	100	0	0	0	**192**
GPT (12,288)	**60 days**	86	24	93	0	93	0	**296**
	120 days	85	53	93	0	93	0	**324**
	240 days	85	86	93	0	93	0	**357**

This table reports the number of times out of 100, that the time series of standard deviations of portfolio returns has a lower mean (statistically significant at 10%) than the given alternatives. Each day a set of 25 companies is randomly selected from the S&P 500 constituents and combined into portfolios using either random weights (RND) or minimum semantic concentration weights (based on the cosine similarities of business descriptions vectorised using three different NLP methods: SF, BERT, GPT). The standard deviation of each portfolio constructed on a given day is calculated assuming the portfolio is held for the subsequent 60, 120 and respectively 240 days.

Table 4. STOXX europe 600: Minimum semantic concentration portfolios: 25 companies.

		RND	SF	BERT	BERT	BERT	GPT	**Wins**
	RW		(16,384)	(768)	(348)	(1024)	(12,288)	**Total**
RND	**60 days**	0	0	0	0	0	0	**0**
	120 days	0	0	0	0	0	0	**0**
	240 days	0	0	0	0	0	0	**0**

Continued

SF (16,384)	**60 days**	93	0	100	100	0	100	**393**
	120 days	93	0	100	100	0	100	**393**
	240 days	94	0	100	100	0	100	**394**
BERT (768)	**60 days**	93	0	0	0	0	0	**93**
	120 days	93	0	0	0	0	0	**93**
	240 days	93	0	0	0	0	0	**93**
BERT (348)	**60 days**	93	0	100	0	0	84	**277**
	120 days	93	0	100	0	0	98	**291**
	240 days	94	0	100	0	0	100	**294**
BERT (1024)	**60 days**	93	0	100	100	0	100	**393**
	120 days	94	0	100	100	0	100	**394**
	240 days	94	0	100	100	0	100	**394**
GPT (12,288)	**60 days**	93	0	100	0	0	0	**193**
	120 days	93	0	100	0	0	0	**193**
	240 days	94	0	100	0	0	0	**194**

This table reports the number of times out of 100, that the time series of standard deviations of portfolio returns has a lower mean (statistically significant at 10%) than the given alternatives. Each day a set of 25 companies is randomly selected from the STOXX Europe 600 constituents and combined into portfolios using either random weights (RND) or minimum semantic concentration weights (based on the cosine similarities of business descriptions vectorised using three different NLP methods: SF, BERT, GPT). The standard deviation of each portfolio constructed on a given day is calculated assuming the portfolio is held for the subsequent 60, 120 and respectively 240 days.

Table 5. S&P 500: Minimum semantic concentration portfolios: 50 companies.

		RND	SF	BERT	BERT	BERT	GPT	**Wins**
	RW		(16,384)	(768)	(348)	(1024)	(12,288)	**Total**
RND	**60 days**	0	0	0	0	0	0	**0**
	120 days	0	0	0	0	0	0	**0**
	240 days	0	0	0	0	0	0	**0**
SF (16,384)	**60 days**	93	0	100	0	100	0	**293**
	120 days	93	0	100	0	100	0	**293**
	240 days	93	0	100	0	100	0	**293**
BERT (768)	**60 days**	93	0	0	0	0	0	**93**
	120 days	92	0	0	0	0	0	**92**
	240 days	93	0	0	0	0	0	**93**

Continued

BERT (348)	**60 days**	93	0	100	0	100	0	**293**
	120 days	93	0	100	0	100	0	**293**
	240 days	93	0	100	0	100	0	**293**
BERT (1024)	**60 days**	93	0	100	0	0	0	**193**
	120 days	93	0	100	0	0	0	**193**
	240 days	93	0	100	0	0	0	**193**
GPT (12,288)	**60 days**	77	0	83	0	83	0	**243**
	120 days	77	0	83	0	83	0	**243**
	240 days	77	3	83	0	83	0	**246**

This table reports the number of times out of 100, that the time series of standard deviations of portfolio returns has a lower mean (statistically significant at 10%) than the given alternatives. Each day a random set of 50 companies is randomly selected from the S&P 500 constituents and combined in portfolios using either random weights (RND) or minimum semantic concentration weights (based on the cosine similarities of business descriptions vectorised using three different NLP methods: SF, BERT, GPT). The standard deviation of each portfolio constructed on a given day is calculated assuming the portfolio is held for the subsequent 60, 120 and respectively 240 days.

Table 6. STOXX europe 600: Minimum semantic concentration portfolios: 50 companies.

		RND	SF	BERT	BERT	BERT	GPT	**Wins**
	RW		(16,384)	(768)	(348)	(1024)	(12,288)	**Total**
RND	**60 days**	0	0	0	0	0	0	**0**
	120 days	0	0	0	0	0	0	**0**
	240 days	0	0	0	0	0	0	**0**
SF (16,384)	**60 days**	100	0	100	100	100	100	**500**
	120 days	92	0	100	100	100	100	**492**
	240 days	94	0	100	100	0	100	**394**
BERT (768)	**60 days**	100	0	0	0	0	0	**100**
	120 days	92	0	0	0	0	0	**92**
	240 days	93	0	0	0	0	0	**93**
BERT (348)	**60 days**	100	0	100	0	0	0	**200**
	120 days	92	0	100	0	0	0	**192**
	240 days	94	0	100	0	100	0	**294**
BERT (1024)	**60 days**	100	0	100	100	0	0	**300**
	120 days	92	0	100	100	0	100	**392**
	240 days	94	0	100	100	0	100	**394**

Continued

GPT (12,288)	**60 days**	100	0	100	100	0	0	**300**
	120 days	92	0	100	54	0	0	**246**
	240 days	94	0	100	0	0	0	**194**

This table reports the number of times out of 100, that the time series of standard deviations of portfolio returns has a lower mean (statistically significant at 10%) than the given alternatives. Each day a random set of 50 companies is randomly selected from the STOXX Europe 600 constituents and combined in portfolios using either random weights (RND) or minimum semantic concentration weights (based on the cosine similarities of business descriptions vectorised using three different NLP methods: SF, BERT, GPT). The standard deviation of each portfolio constructed on a given day is calculated assuming the portfolio is held for the subsequent 60, 120 and respectively 240 days.

5. Concluding Remarks

As academics and practitioners alike recognise the value of including unstructured (text) data in their analyses, they are confronted with an ever increasing choice of NLP methods that attempt to "understand" natural language. The more sophisticated methods (such as the ones reviewed in this paper) tend to outperform the simpler ("bag-of-words") approaches, but come at the disadvantage of being virtual "blackboxes" for the end-user. In these conditions, choosing the appropriate method depends on how well it performs a specific task, rather than on the merits of its algorithmic design. This paper compares the performance of three NLP methods in a test that is central to modern finance, namely the ability to diversify risk in a portfolio of assets and thereby aims to provide guidance to users of text analysis, in the process of choosing a particular NLP method. By converting business descriptions of S&P 500 and respectively STOXX Europe 600 constituent companies into numerical representations of meaning and computing their similarity, so called minimum semantic concentration portfolios can be constructed by minimising the semantic similarity across companies (akin to the minimum variance portfolios standard in the literature). A "horse race" is run across three NLP methods (two large language models: BERT and GPT as well as an alternative approach, semantic fingerprinting) and the results show that all the NLP methods result in portfolios with significantly lower volatility than a portfolio with randomly chosen weights. While all NLP methods are able to extract meaningful information from the input data, none of the methods emerges as a clear winner. Given that large language models come at a significant cost not only financially but also in environmental terms due to their resource intensive computational needs, cost effectiveness might become and important factor in choosing a particular NLP method. As an alternative to the pre-trained large language models (such as BERT and GPT), semantic fingerprinting (SF), currently an "off-the-shelf" AI solution, provides an easy to implement, albeit not-customisable solution, which performs robustly in a consistent fashion. This could be optimal for an end-user who prefers a reliable choice and does not wish to commit resources to further

training and/or fine-tuning models. In contrast, the performance of the large language models is more volatile even across tasks that are extremely similar, which suggests that the end-user whose NLP choice is a large language model should be prepared to invest into extensive comparisons and further training, with the aim of obtaining superior results.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Alexeev, V. V., & Tapon, F. (2013). Equity Portfolio Diversification: How Many Stocks Are Enough? Evidence from Five Developed Markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2182295

Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics, 15,* 659-688. https://doi.org/10.1146/annurev-economics-082222-074352

Chen, L., Zaharia, M., & Zou, J. (2023). *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*. Working Paper, Under Review as a Conference Paper at ICLR 2024.

De Sousa Webber, F. (2016). *Semantic Folding Theory and Its Applications in Semantic Fingerprinting*. White Paper, arXiv: 1511.08855.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186). Association for Computational Linguistics.

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E. et al. (2022). Measuring the Carbon Intensity of AI in Cloud Instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1877-1894). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533234

Hawkins, J. (2021). *A Thousand Brains: A New Theory of Intelligence*. Hachette.

Hawkins, J., Ahmad, S., & Cui, Y. (2017). A Theory of How Columns in the Neocortex Enable Learning the Structure of the World. *Frontiers in Neural Circuits, 11,* Article 81. https://doi.org/10.3389/fncir.2017.00081

Ibriyamova, F., Kogan, S., Salganik-Shoshan, G., & Stolin, D. (2017). Using Semantic Fingerprinting in Finance. *Applied Economics, 49,* 2719-2735. https://doi.org/10.1080/00036846.2016.1245844

Ibriyamova, F., Kogan, S., Salganik-Shoshan, G., & Stolin, D. (2019). Predicting Stock Return Correlations with Brief Company Descriptions. *Applied Economics, 51,* 88-102. https://doi.org/10.1080/00036846.2018.1494377

Pungulescu, C. (2022a). Bilateral Home Bias: A New Measure of Proximity. *Journal of Neuroscience, Psychology, and Economics, 15,* 163-177. https://doi.org/10.1037/npe0000162

Pungulescu, C. (2022b). Using Textual Analysis to Diversify Portfolios. *The Economics and Finance Letters, 9,* 87-98. https://doi.org/10.18488/29.v9i1.3028

Pungulescu, C. (2024). *Predicting Return Correlations in European Stocks Using NLP*. Working Paper.

Pungulescu, C., & Stolin, D. (2023). Measuring Document Similarity: A Comparative Analysis of NLP Methods in Finance. *Mendeley Data*. https://doi.org/10.17632/kmb89v8yhz.1