**The Economics and Finance Letters**
*2022 Vol. 9, No. 1, pp. 87-98.*
*ISSN(e): 2312-430X*
*ISSN(p): 2312-6310*
*DOI: 10.18488/29.v9i1.3028*


# USING TEXTUAL ANALYSIS TO DIVERSIFY PORTFOLIOS

**Crina Pungulescu**
*Institute for Economic Forecasting, National Institute of Economic Research and Romanian Academy, Bucharest, Romania.*
*Email: crina_pungulescu@ipe.ro*

**Article History**
Received: 7 March 2022
Revised: 10 May 2022
Accepted: 26 May 2022
Published: 14 June 2022

**Keywords**
Textual analysis
Stock returns correlations
Portfolio performance
Mean Variance investing
Natural language understanding.

**JEL Classification:**
G11; G15.

## *ABSTRACT*

Semantic fingerprinting is a leading AI solution that combines recent developments from cognitive neuroscience and psycholinguistics to analyze text with human-level accuracy. As an efficient method of quantifying text, it has already found its application in finance where the semantic fingerprints of company descriptions have been shown to successfully predict stock return correlations of Dow Jones Industrial Average (DJIA) constituents. By extension, it has been suggested that diversified portfolios could be constructed to exploit the fundamental (dis)similarity between companies' core activities (measured by the semantic overlap of company descriptions). This paper follows the performance of two portfolios made of the same DJIA constituent companies: the "minimum semantic concentration" portfolio (constructed with text-based portfolio weights) and the traditional "minimum variance" portfolio, over a time span of 16 years including two high volatility events: the 2007 – 2009 financial crisis and the COVID pandemic. The results confirm that textual analysis using semantic fingerprinting is consistently successful in predicting stock return correlations and is valuable as a portfolio selection criterion. However, in times of high market volatility the fundamental information given by the companies' core activities, while still relevant, might carry less weight.

**Contribution/Originality:** This study is the first to test empirically the performance of a portfolio that minimizes the semantic concentration (or similarity) of its constituent companies' descriptions. The "minimum semantic concentration" (MSC) portfolio consistently outperforms the traditional "minimum variance portfolio" (MV) over a 16 year period.

## 1. INTRODUCTION

Driven by technological advances that allow for the quantification of qualitative information, textual analysis is increasingly becoming part of the essential toolset in empirical finance. In their earlier survey of textual analysis in finance and accounting, Loughran and McDonald (2016) emphasise both the wealth of information that can be mined from available text (e.g. Securities and Exchange Commission filings for U.S. companies or other forms of company descriptions, press releases, news and social media), and the unavoidable subjectivity and limitations of pre-defined "word lists" and related methodology that rely intensively on human decisions. Whether machine learning methodologies will reach the necessary level of sophistication to perform this work with human-level accuracy was merely an open question at the time. In a subsequent review of the field, Gentzkow, Kelly, and Taddy (2019) are able to answer in the affirmative and suggest that the great advances in machine learning will be successfully leveraged in valuable economic research. Machine learning solutions are attractive due to their ability

to analyze large text databases and to reduce dramatically researchers' subjectivity, once the algorithm is set up. Their quality can only be measured by the usefulness of their contribution to testing economic hypotheses.

This paper uses one such algorithm that appeals in terms of its intuitive design as well as ease of implementation and interpretation. Semantic fingerprinting is a machine learning algorithm designed by *Cortical.io*[1], whose technology is grounded in neuroscience and emulates the way the human brain works in processing information and creating associations between various concepts. The technology has been counted along with *IBM*'s and *Amazon*'s solutions among the top 10 keyword extraction APIs[2], but its most important contribution is the unique ability to convert efficiently and automatically any text into a numerical representation of its meaning.

*Cortical.io*'s AI solution to quantifying text involves the so-called *Retina engine*[3], that has been trained on virtually the entire collection of Wikipedia articles of 2014 and distilled all the information into a semantic map organised as a 128×128 matrix of 16,384 learned "contexts" or topics that are clustered together based on the associations of their meanings (as "learned" from "reading" Wikipedia in its entirety). The semantic fingerprint[4] of any text is a selection of (at most) 984 of the 16,384 predefined "contexts" that are most associated with the text's meaning. A detailed description of semantic folding[5], and its application to semantic fingerprinting, is provided in *Cortical.io*'s White Paper (De Sousa Webber, 2015).

This method of text quantification provides an immediate solution for the semantic comparison of texts and as such, becomes a viable methodology for identifying peer companies where the similarity criterion depends on a certain textual description of the company. Since the semantic fingerprint of any chosen text can be represented as a binary vector of 16,384 positions (with 1s for the 984 "contexts" describing any given text), the cosine similarity of the two binary vectors provides a direct and intuitive measure of semantic overlap. For all but the shortest texts, the *Retina engine* returns a semantic fingerprint made of the maximum number of 984 "contexts" and, in this case, the cosine similarity becomes the fraction of shared "contexts" between the two fingerprints (see Pungulescu (2020), for a detailed description and examples).

The semantic fingerprinting technology was introduced to finance by Ibriyamova, Kogan, Salganik-Shoshan, and Stolin (2017) who show that using the cosine similarity between the semantic fingerprints of company descriptions outperforms "word lists" in predicting stock return correlations for the 30 DJIA constituents (as of the end of 2012). The semantic fingerprints used by Ibriyamova et al. (2017) are based on extensive descriptions available in the 10-K reports required by the Securities and Exchange Commission (S.E.C.). Since such detailed descriptions are only available for U.S. companies, textual analysis of international stocks appears to be at an informational disadvantage. Ibriyamova, Kogan, Salganik-Shoshan, and Stolin (2019) remedy that by showing that the predictive power of globally available (shorter) company descriptions is even higher for stock return correlations than that of the longer 10-K filings used previously.

Given the role played by stock returns correlations in financial diversification and, therefore, portfolio selection, Ibriyamova et al. (2019) suggest that a diversified portfolio can be constructed based solely on company descriptions, since similar companies tend to behave similarly. They further show theoretically that the cosine similarity portfolio weights are the solution to minimizing semantic concentration (or semantic similarity), subject to the budget constraint, in a process that is mathematically identical to the identification of the "minimum variance" portfolio. The resulting "minimum semantic concentration" (or, equivalently, "maximum semantic diversification") portfolio

[1] https://www.cortical.io/

[2] https://www.edenai.co/post/top-10-keyword-extraction-api

[3] https://www.cortical.io/technology/retina-engine/

[4] Short videos illustrating semantic fingerprinting are available at https://www.cortical.io/resources/#videos

[5] *Cortical.io*'s Semantic Folding complements *Numenta, Inc.*'s Hierarchical Temporal Memory (HTM) theory which models the algorithmic functionalities of the human neocortex (Hawkins, Ahmad, Purdy, & Lavin, 2019).

provides, therefore, a natural alternative to the minimum variance investing rule that gained popularity among practitioners as evidence mounted that the "minimum variance" portfolio outperforms the market-weighted portfolio for U.S. as well as for global stocks (see Scherer (2011), and references therein).

This paper is the first to explore the empirical performance of the "minimum semantic concentration" (MSC) portfolio proposed by Ibriyamova et al. (2019) and it does that by comparing the performance of two portfolios: the "minimum semantic concentration" (MSC) and the traditional "minimum variance" weights (MV) portfolio, over a span of 16 years. The time dimension is particularly relevant, since cosine similarities differ from stock return correlations in a fundamental way: similarities based on company descriptions can only change if and when a company changes its core activities (and, therefore, its description changes), whereas stock return correlations are notoriously unstable over time with higher return correlations linked to periods of high market volatility (Loretan & English, 2000). Ibriyamova et al. (2017); Ibriyamova et al. (2019) proved the validity of semantic fingerprinting of 2013 company descriptions in predicting stock return correlations in 2014, a period of relatively low market volatility. The long timespan of this paper covers several high volatility episodes (including the 2007 – 2009 financial crisis and the 2020 – 2022 COVID pandemic), which makes it possible to investigate how the impact of cosine similarity on stock return correlations changes over time. In this context, it becomes an empirical question whether computing portfolio weights based on the cosine similarity of semantic fingerprints of company descriptions is equally valid as a portfolio selection criterion in times of high or low volatility.

The results support the view that cosine similarity of company descriptions is consistently successful at predicting stock return correlations. In over 4,000 rolling windows regressions, the coefficients of the cosine similarity variable are statistically significant at 10% in 75% of the models and statistical significance tends to be higher in periods of low to moderate volatility. A similar pattern is noticeable when comparing the performance of the MSC portfolio with the performance of the MV portfolio. The MSC portfolio outperforms the MV portfolio (in terms of mean to standard deviation ratio), both in the periods with above and below average volatility, but in times of high volatility the MSC portfolio performs better than the MV portfolio only 57% of time versus 70% of the time when volatility is below average.

The remainder of this paper is organized as follows. Section 2 places semantic fingerprinting within the context of the textual analysis techniques currently used in applications to finance, section 3 reviews the data and the methodology, while section 4 presents the results of the empirical analysis and section 5 presents the conclusions.

## 2. LITERATURE REVIEW

With great illustrative force, Pratt (2015) likened the power of "deep learning algorithms" that train robots on extensive datasets and "cloud robotics" that allow the sharing of learning among robots to the Cambrian explosion (when, more than 500 million years ago, a great diversity of organisms emerged over a relatively short period). The semantic fingerprinting technology used in this paper is one of the many "organisms" that emerged during this potentially explosive stage in the evolution of AI.

The pioneering work of Ibriyamova et al. (2017); Ibriyamova et al. (2019), who introduce this technology to finance, and the extension provided by this paper are motivated by the availability of this fast, reliable and intuitive method of mining text for relevant, easily quantifiable information. In the quest for an AI solution that is able to successfully compete with humans in the ability to perceive nuance and handle complexity, this technology is an especially interesting candidate since it emulates the way the human brain learns concepts based on their association and stores related concepts closely together. Granting, of course, that humans process language better than any computer to date, we may think of the *Retina engine* as an AI "brain" that has learned everything it knows from reading virtually the entire Wikipedia. In an accurate parallel to the human brain, the *Retina engine* created its own semantic map of associated meanings or "contexts" and on receiving any new piece of information it 'fits' it on the pre-existing semantic map based on the ways it relates to the previously learned "contexts". Topology is important

in the semantic space with non-related "contexts" situated in distant positions on the map, which allows for a clear separation between homonyms (i.e. investment "bank" vs. river "bank"). In consequence, with its focus on decomposing text into meaning with near-human accuracy, semantic fingerprinting improves on the traditional (necessarily reductive) "word list" methods, traditionally used in textual analysis, by correctly identifying and disambiguating among different contexts and by reducing subjectivity to a single researcher choice: the relevant text to be fingerprinted.

Semantic fingerprinting is only one of several competing methodologies and it may be compared, for instance, with the Explicit Semantic Analysis (ESA) of Egozi, Markovitch, and Gabrilovich (2011) which also relies on Wikipedia to establish semantic context and uses concept-based features with the aim of identifying similarity between texts that describe the same notion using different terms. However, their methodology acknowledges a limitation when it comes to longer texts. Semantic fingerprinting also shares the goals of the simpler *Word2Vec* model (Mikolov, Chen, Corrado, & Dean, 2013) and *Stanford*'s *GloVe* model (Pennington, Socher, & Manning, 2014) and their "deep learning" counterparts: Embeddings from Language Models (ELMo) representations proposed by Peters et al. (2018) and Bidirectional Encoder Representations from Transformers (BERT) of Devlin, Chang, Lee, and Toutanova (2018).

While the inner workings of all these technical solutions are relatively opaque for the typical empirical researcher, it should be noted that their essential aim is to provide an input for subsequent empirical analysis in a manner that is replicable and lends itself to applications in testing a wide range of hypotheses. Ultimately, any method's value derives from its ability to provide informative variables to subsequent empirical research.

Once quantified text is available, one of its important applications in finance (and the one relevant to this paper) is to obtain a company similarity measure. As the quantified text is typically expressed as a vector, using the cosine similarity has already become the established approach, mainly due to its simplicity and ease of interpretation as the angle between two word vectors on a unit sphere. Brown and Tucker (2011) apply it to Management Discussion and Analysis (MD&A) disclosures, Lang and Stice-Lawrence (2015) use annual reports to compare firms based on the cosine similarity of vectors comparing the relative word frequencies across documents and Hoberg and Phillips (2016) compare vectors of unique words appearing in each company's S.E.C. filings to create an alternative, time-varying set of text-based industry classifications. Bushman, Chen, and Williams (2017) use both S.E.C. filings and Management Discussion and Analysis (MD&A) disclosures to identify the clusters of banks that have similar exposures to downside tail risk. Text-based similarity measures have also been successfully employed in comparing technological profiles of companies based on their patent profiles (Arts, Cassiman, & Gomez, 2018; Kelly, Papanikolaou, Seru, & Taddy, 2021; Testoni, 2021). The list of potential applications is virtually limitless, which makes finding an accurate and accessible technology even more desirable.

In their application to predicting stock return correlations, Ibriyamova et al. (2017) run a 'horse race' among several similarity measures, alongside the one provided by semantic fingerprinting, and show that the latter has a significantly higher predictive power than the Hoberg & Philips measure (Hoberg & Phillips, 2016). At the same time, the similarity measure based on semantic fingerprinting compares well with other candidates for identifying peer companies, based on similar patterns in "analyst coverage" (Kaustia & Rantala, 2021) or "Internet searches" (Lee, Ma, & Wang, 2015). The latter two methods proved similarly successful in predicting stock return correlations but are constrained by design to fit a smaller set of research questions. Ease of implementation, computational efficiency and virtually unlimited versatility in application might give semantic fingerprinting an extra edge.

Surveys of recent trends in textual analysis in finance (Loughran & McDonald, 2020) and accounting (Bochkay, Brown, Leone, & Tucker, 2022) suggest that reliance on refined machine learning approaches is on the rise, a trend that is justified by the advantages they bring. Investment in state-of-the-art textual analysis technologies is set to become part of the essential toolset of any researcher.

## 3. DATA AND METHODOLOGY

Daily price data is retrieved from Bloomberg for the 3.01.2005 – 14.01.2022 period for the same 30 DJIA constituents[6] used by Ibriyamova et al. (2017); Ibriyamova et al. (2019) and continuously compounded daily returns are calculated as $r_t = \ln(p_t) - \ln(p_{t-1})$, where $p_t$ is the closing price (adjusted for dividends) on day $t$ for each company. Realized daily volatility is computed as the square root of the sum of daily squared returns for the 30 companies and pairwise stock return correlations are calculated for 60–day rolling windows, a practical choice for forecasting correlations (see Jeon and McCurdy (2017), for a detailed discussion).

Short company descriptions were retrieved from Thomson Reuters and their semantic fingerprints were obtained using the *Cortical.io* technology[7]. Appendix A provides examples of descriptions for four DJIA constituents: Disney, Pfizer, Bank of America and JPMorgan Chase. These four companies were chosen to illustrate the range of cosine similarities in the sample: cosine similarity is lowest, 0.18, for the Disney and Pfizer pair and highest, 0.65, for Bank of America and JPMorgan Chase. For comparison, the average cosine similarity among the 30 DJIA constituent companies is 0.39. Figure 1 shows the histogram of the cosine similarities of the semantic fingerprints corresponding to the resulting 435 company pairs, while Figures 2 and 3 provide an illustration of the semantic space as a 128 × 128 matrix of 16,384 pre-defined "contexts" and an intuitive way to visualize the concept of semantic overlap for the Bank of America/JPMorgan Chase and the Disney/Pfizer pairs respectively.

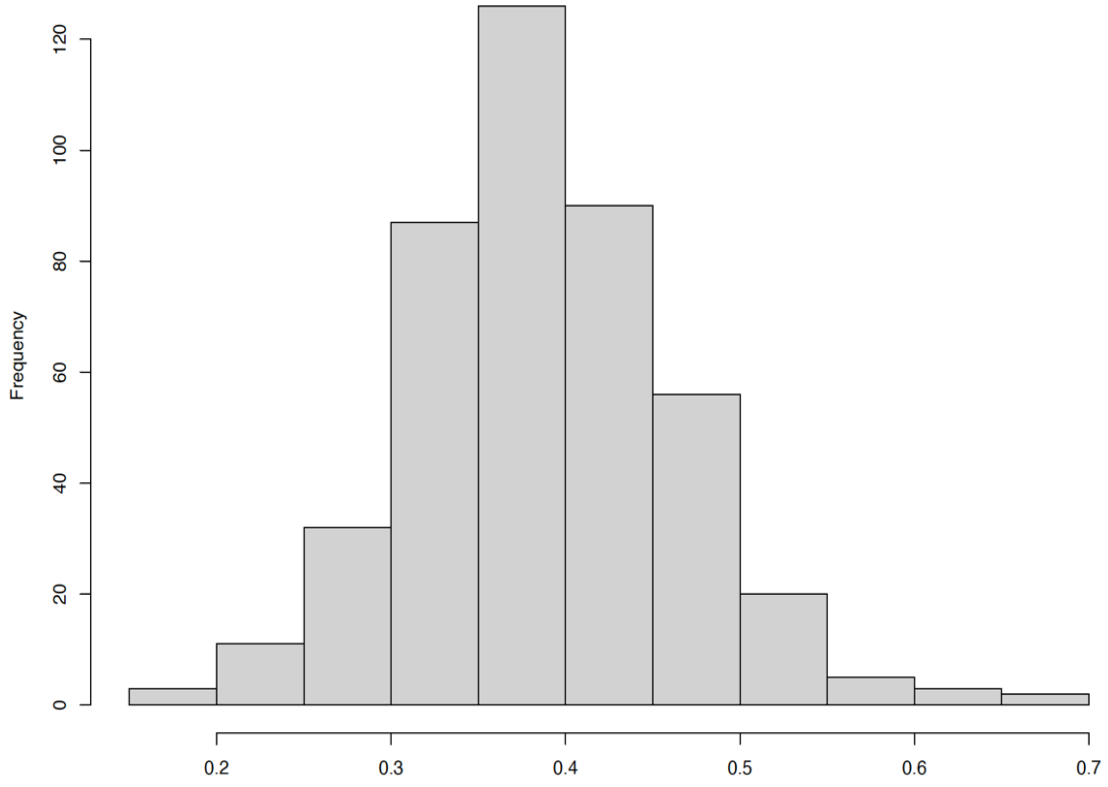


**Figure 1.** Histogram of cosine similarities.
**Note:** This figure plots the time-varying average pairwise correlations of stock returns against the realized volatility.

[6] For United Technologies Corporation, data is available only until 3.04.2020, the date when it underwent a merger of equals with Raytheon Technologies Corporation.

[7] The *Retina engine* can be accessed at https://tinyurl.com/retinaengine. Note that, consistently with a convention used by many programming languages, *Cortical.io* numbers the positions starting from 0 and not from 1.

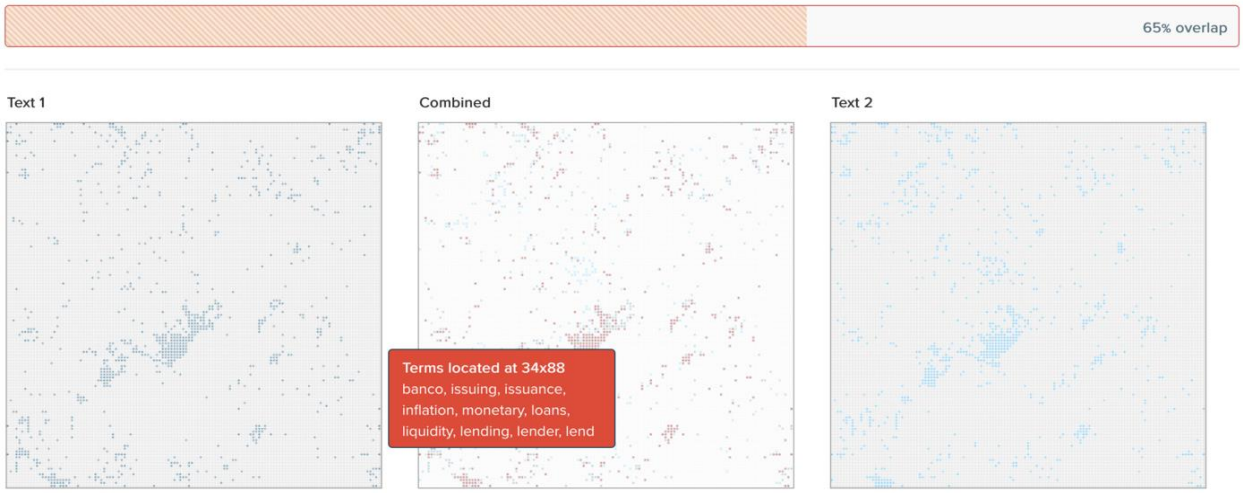


**Figure 2.** High semantic overlap: Bank of America and JPMorgan Chase.

**Note:** This figure is provided by the similarity explorer available at https://www.cortical.io/freetools/compare-text/. It illustrates the highest example of semantic overlap among the DJIA constituents. Bank of America and JPMorgan Chase share 65% of the "contexts" that represent their semantic fingerprints in the 128×128 semantic universe. The texts fingerprinted are the companies' descriptions from Thomson Reuters and the 984 positions that make each text's fingerprint are highlighted in blue. The "contexts" shared by the two fingerprints are highlighted in red in the center panel. One such "context" situated in the position 34×88 in the semantic space identifies meanings related to the business both companies are operating in (such as banking, money and lending).

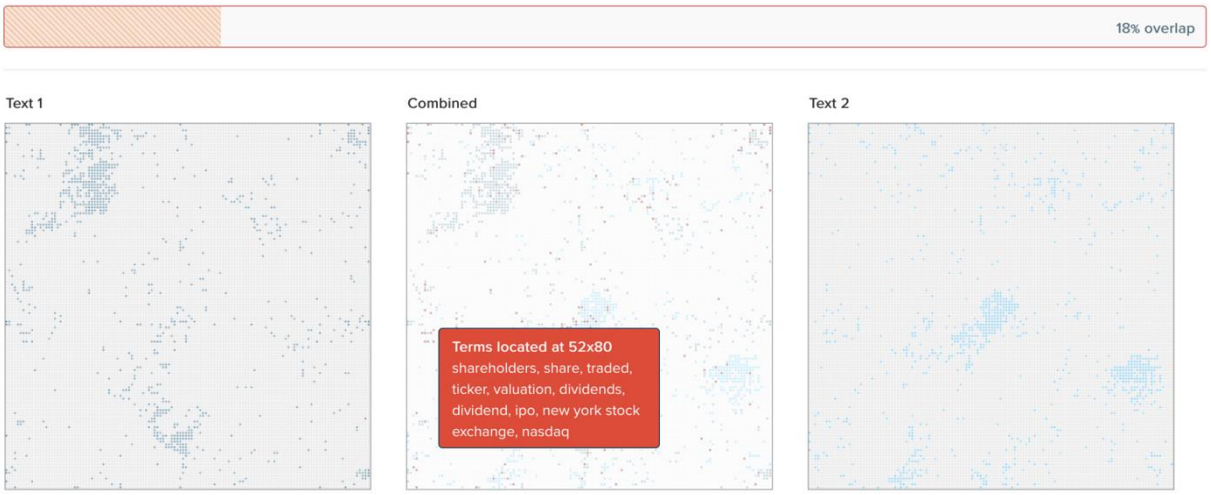


**Figure 3.** Low semantic overlap: Disney and Pfizer.

**Note:** This figure is provided by the similarity explorer available at https://www.cortical.io/freetools/compare-text/ . It illustrates the lowest example of semantic overlap among the DJIA constituents. Disney and Pfizer share 18% of the "contexts" that represent their semantic fingerprints in the 128×128 semantic universe. The texts fingerprinted are the companies' descriptions from Thomson Reuters and the 984 positions that make each text's fingerprint are highlighted in blue. The "contexts" shared by the two fingerprints are highlighted in red in the center panel. One such "context" situated in the position 52×80 in the semantic space identifies meanings related to publicly traded companies (such as shareholders, dividends and stock exchanges).

Each figure consists of three panels representing the matrix of semantic space where the fingerprints of the two companies compared semantically appear as 984 highlighted "contexts" in the left and right panels (in shades of blue). The middle panel allows for the two fingerprints to "overlap" in the semantic space and the "contexts" shared by both fingerprints are highlighted in red. Moreover, *Cortical.io*'s similarity explorer tool[8] allows identification of each "context" belonging to one or common to both fingerprint(s), simply by positioning the cursor over any highlighted position in the semantic space. Semantic fingerprinting correctly identifies the high similarity between two companies operating in the same field (for Bank of America and JPMorgan Chase) as well as the fact that an entertainment and a pharmaceutical company have little in common apart from the fact that they are both publicly traded companies. It is worth noting that topology is relevant with related "contexts" positioned adjacently on the map. The two investment banks not only share a large proportion of "contexts" but evidently share entire clusters

[8] Available at https://www.cortical.io/freetools/compare-text/

of related "contexts", whereas the two companies operating in unrelated fields share only a few and only isolated "contexts". With the aid of geographical directions, one could say that the main cluster of "contexts" for Disney lies in the North West area of the semantic map, where "contexts" related to "network", "cable", "broadcasting" etc. are stored, followed by a second cluster positioned in the South of the map, which stores "contexts" related to "film", "cinematography", "picture" and so on. Not surprisingly, the main cluster for Pfizer (located centrally on the semantic map) relates to topics like "healthcare", "prescriptions" and "pharmacy" followed by another sizeable cluster in the South East area, which stores "contexts" related to various organs ("liver", "lungs" etc.) and related diseases. Among the few shared "contexts" we find topics related to "business", "products" and "shareholders" but also topics involving "children" (in the context of paediatrics for Pfizer).

The methodological choices of this paper aim to answer three questions.

Firstly, this paper asks whether the impact of cosine similarity on stock returns correlations changes over time. To answer this question, the following equation from Ibriyamova et al. (2019) is re-estimated over a number of 4,171 (60−day) rolling windows:

$$z_{i,j,t} = \beta_0 + \beta_1 z_{i,j,t-1} + \beta_2 \sigma_{i,t-1}\sigma_{j,t-1} + \beta_3 s_{i,j} + \epsilon_{i,j,t}, \quad (1)$$

Where $z_{i,j,t} = 0.5 \ln\left(\frac{1+\rho}{1-\rho}\right)$ is the Fisher transformation of the correlation between the stock returns of company $i$ and company $j$ in period $t$; $\sigma_{i,t}$ is the standard deviation of the stock return of company $i$ in period $t$; $\sigma_{j,t}$ is the standard deviation of the stock return of company $j$ in period $t$ and $s_{ij}$ is the cosine similarity between the semantic fingerprints of the company descriptions for companies $i$ and $j$.

The second question of this paper is to investigate the performance of the "minimum semantic concentration" (MSC) portfolio constructed based on the theoretical arguments of Ibriyamova et al. (2019). Ibriyamova et al. (2019) define the semantic concentration of a portfolio as the sum of the squared elements of the vector obtained by multiplying the matrix of binary semantic fingerprints of each constituent asset and the vector of portfolio weights (similar to the way the Herfindahl-Hirschman Index measures market concentration). Minimizing the semantic concentration of a given portfolio subject to the budget constraint is mathematically identical to the problem of identifying the global minimum variance portfolio and results in the same analytical solution: the minimum semantic concentration (MSC) portfolio weights are given by $\frac{(S^T S)^{-1}1}{1^T (S^T S)^{-1}1}$, where $S^T S$, the matrix of cosine similarities, takes on the role that the variance-covariance matrix, $\Sigma$, plays in finding the minimum variance (MV) portfolio weights. Comparing the performance of the minimum semantic concentration (MSC) and the minimum variance (MV) portfolio in terms of their mean to standard deviation ratios, becomes a natural empirical question.

Finally, as the two portfolios are followed over 4,171 (60−day) rolling windows covering a 16-year period and including the 2007 − 2009 financial crises and the 2020 − 2022 COVID pandemic, this paper asks how the performances of the two portfolios compare in high volatility environments, which are defined as the periods when the realized volatility is above the third quartile.

## 4. EMPIRICAL ANALYSIS

As expected, there is an evident link between daily realized volatility and average stock return correlation, as Figure 4 shows.

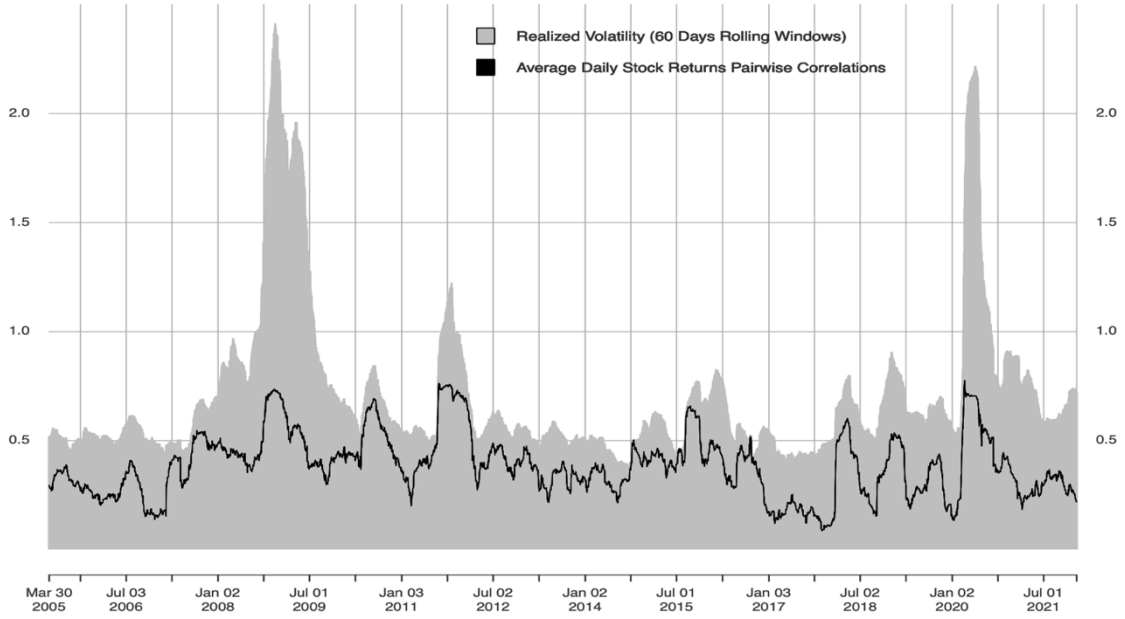


**Figure 4.** Realized volatility & stock returns correlations.
**Note:** This figure plots the time-varying average pairwise correlations of stock returns against the realized volatility.

The pairwise correlation of the stock returns and realized volatility is high, at 0.63 (statistically significant at any level of significance, with a $t$–statistic of 53.7). In contrast, the cosine similarity based on semantic fingerprints (in effect, the fraction of semantic "contexts" that are shared by two companies) remains constant throughout the sample. This motivates the dynamic component of the analysis conducted in this paper, given that the results of Ibriyamova et al. (2017); Ibriyamova et al. (2019) were obtained in a period of relatively moderate volatility and the predictive power of the cosine similarity on returns correlations is apt to vary with market conditions.

It is unsurprising that the coefficients of the cosine similarity ($\beta_3$ in the Equation 1) vary considerably over time, as Figure 5 illustrates.

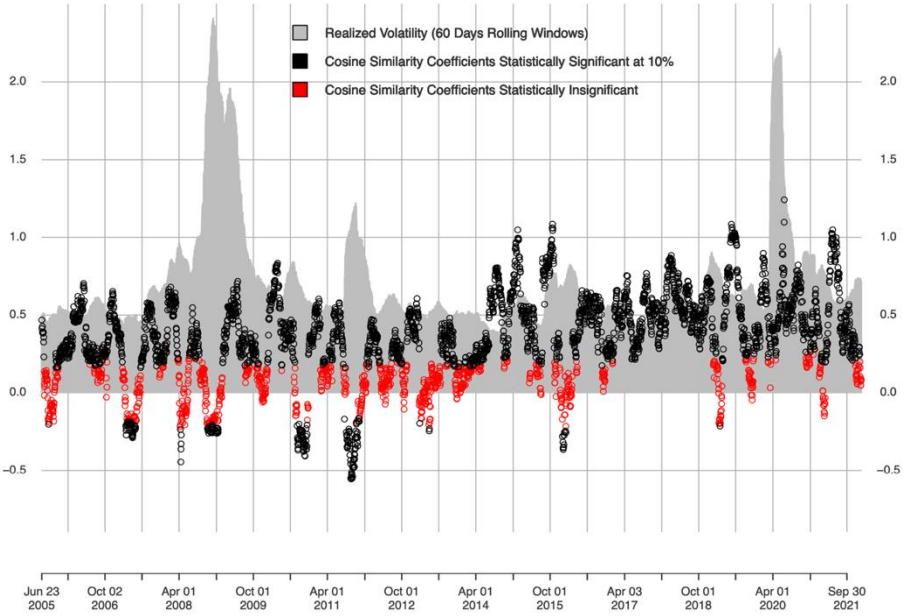


**Figure 5.** Time-varying cosine similarity coefficients.
**Note:** This figure shows the time-varying cosine similarity coefficients (when statistically significant at 10% in black and when statistically insignificant in red). The realized volatility is shown in background in grey.

However, the coefficients remain positive and statistically significant at 10% in 3.131 estimations (75% of the total), illustrating the consistency of cosine similarity as a predictor of stock return correlations over a long time span. As the time varying coefficients of cosine similarity are plotted against the background of realized volatility in Figure 5, there is a suggestion that the impact of cosine similarity, a stable measure of fundamental characteristics shared by two companies, on the (dynamic) stock returns correlations becomes less manifest in times of increased volatility. Indeed, the pairwise correlation between realized volatility and the coefficients of cosine similarity is negative, specifically −0.15, and statistically significant at any level of significance, with a $t$−statistic of −9.7. When market volatility increases, the impact of the cosine similarity of the companies' semantic fingerprints on stock returns correlations diminishes.

Next, the MSC and MV portfolios were compared in terms of the ratio of mean returns to their standard deviation computed for each of the (60−day) rolling windows. There is a clear promise in semantic diversification: the MSC portfolio performed best in 2.753 periods (66% of the time), while the MV portfolio performed best 1.417 times. Figure 6 shows the periods when the MSC portfolio performed best (in black) and the periods when the MV portfolio performed best (in red) over the background of the daily realized volatility.

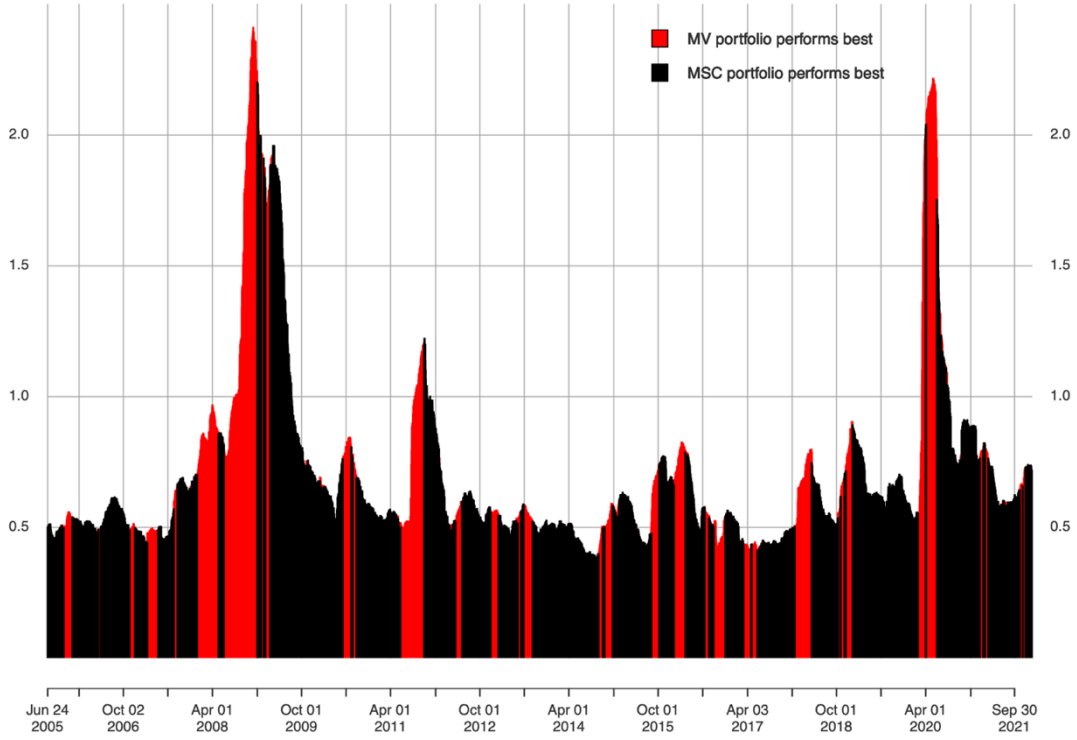


**Figure 6.** Realized volatility (60-day Rolling Windows).

**Note:** This figure shows the realized volatility with the periods when the MV portfolio performed better than the MSC portfolio shown in red and the periods when the MSC portfolio performed better shown in black.

The figure also suggests that the MSC portfolio performed better in periods of lower volatility. Restricting the sample to periods of high volatility (defined as periods where the realized volatility is above the third quartile) showed that the MSC performed better than the MV portfolio only 57% of the time (versus 70% of the time when the sample is restricted to periods when the realized volatility is below average, for instance).

Over the entire sample, the daily[9] realized volatility amounted to an average of 0.64 in the days when the MSC portfolio performed best versus 0.84 in the days when the MV portfolio performed best. The difference is statistically significant at all levels of significance (with a $t$−statistic of 17.8 and an ANOVA $F$−statistic of 318.3).

[9] The daily realized volatility is the volatility on the last day of the rolling window used to assess portfolio performance.

These results show for the first time that using textual analysis through semantic fingerprinting to construct portfolios has the potential of improving performance. They also suggest that the prospective benefits are dependent on the state of market volatility: in periods of high volatility the minimum variance portfolio has a higher chance of performing better.

## 5. CONCLUDING REMARKS

This paper extends the work pioneered by Ibriyamova et al. (2017); Ibriyamova et al. (2019) who introduced the semantic fingerprinting technology to textual analysis in finance, specifically applied to predicting stock return correlations by adding a time dimension to the analysis. Moreover, the minimum semantic concentration (MSC) criterion suggested by Ibriyamova et al. (2019) was used for portfolio constructing for the first time in empirical analysis. The performance of the resulting minimum semantic concentration (MSC) portfolio was compared with that of the global minimum variance (MV) portfolio, an increasingly popular choice among practitioners.

Adding a time dimension to previous analysis over a time span of 16 years (covering two high volatility events: the financial crisis of 2007 – 2009 and the COVID pandemic of 2000 – 2022) reinforces the validity of semantic fingerprinting, both as a predictor of stock return correlations and as a portfolio selection criterion. The MSC portfolio performed better than the MV portfolio 66% of the time, over the entire sample, but only 57% of the time when the sample is restricted to periods of high volatility.

This paper makes several contributions to the field and aims to provide useful insights to academics and practitioners alike. The first is confirming the validity and the ease of implementation of a technology only recently introduced to finance. Given the increasing importance of refined textual analysis and the wide range of potential applications, this is an opportune moment for any researcher to add an easily implementable and highly accurate tool for quantifying text to their assets. Secondly, this paper proves empirically that textual analysis can provide valuable insights for portfolio selection. The “minimum variance" investment rule (widely used by practitioners) can now be matched or even surpassed by a reliable “minimum semantic concentration" investment rule. One of the advantages is the stability of the portfolio weights over time (since they only change when the company changes its core activities and therefore its description), which makes the cost of rebalancing the portfolio virtually negligible. Finally, it provides a way to design and test many other applications of textual analysis to portfolio selection (where the sources of quantifiable text are not limited to various forms of company descriptions but may include relevant news, forum discussions, tweets etc.).

**Funding:** This research is supported by the Ministry of National Education, CNCS - UEFISCDI (Grant number: PN-II-ID-PCE-2012-4-0631).
**Competing Interests:** The author declares that there are no conflicts of interests regarding the publication of this paper.

## REFERENCES

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, *39*(1), 62-84.Available at: https://doi.org/10.1002/smj.2699.

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2022). Textual analysis in accounting: What's next? *Available at SSRN*, 83.Available at: https://doi.org/10.2139/ssrn.4029950.

Brown, S. V., & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, *49*(2), 309-346.Available at: https://doi.org/10.1111/j.1475-679x.2011.00402.x.

Bushman, R. M., Chen, J. V., & Williams, C. D. (2017). Informativeness and timeliness of text similarity measures for predicting banks' tail comovement. *SSRN Electronic Journal*, 48.Available at: https://doi.org/10.2139/ssrn.2983315.

De Sousa Webber, F. (2015). Semantic folding theory and its application in semantic fingerprinting. Cortical.io White Paper. Retrieved from: https://www.cortical.io/static/downloads/semantic-folding-theory-white-paper.pdf.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Electronic Journal*.Available at: https://doi.org/10.48550/arXiv.1810.04805.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 1-34.Available at: https://doi.org/10.1145/1961209.1961211.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-574.

Hawkins, J., Ahmad, S., Purdy, S., & Lavin, A. (2019). Biological and machine intelligence (BAMI), Initial online release 0.4. Retrieved from: https://numenta.com/resources/biological-and-machine-intelligence/.

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423-1465.Available at: https://doi.org/10.1086/688176.

Ibriyamova, F., Kogan, S., Salganik-Shoshan, G., & Stolin, D. (2017). Using semantic fingerprinting in finance. *Applied Economics*, *49*(28), 2719-2735.Available at: 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*(1), 88-102.

Jeon, Y., & McCurdy, T. H. (2017). Time-varying window length for correlation forecasts. *Econometrics*, *5*(4), 1-29.Available at: https://doi.org/10.3390/econometrics5040054.

Kaustia, M., & Rantala, V. (2021). Common analysts: method for defining peer firms. *Journal of Financial and Quantitative Analysis*, *56*(5), 1505-1536.Available at: https://doi.org/10.1017/s0022109020000514.

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, *3*(3), 303-320.Available at: https://doi.org/10.1257/aeri.20190499.

Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, *60*(2-3), 110-135.Available at: https://doi.org/10.1016/j.jacceco.2015.09.002.

Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410-431.Available at: https://doi.org/10.1016/j.jfineco.2015.02.003.

Loretan, M., & English, W. B. (2000). III. Special feature: Evaluating changes in correlations during periods of high market volatility. *BIS Quarterly Review*, *2*, 29-36.Available at: https://doi.org/10.2139/ssrn.231857.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187-1230.Available at: https://doi.org/10.1111/1475-679x.12123.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, *12*, 357-375.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Electronic Journal*.Available at: https://doi.org/10.48550/arXiv.1301.3781.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv Electronic Journal*.Available at: https://doi.org/10.48550/arXiv.1802.05365.

Pratt, G. A. (2015). Is a Cambrian explosion coming for robotics? *Journal of Economic Perspectives*, *29*(3), 51-60.Available at: https://doi.org/10.1257/jep.29.3.51.

Pungulescu, C. (2020). Bilateral home bias: A new measure of proximity. *Available at SSRN 2697490*, 27.Available at: https://doi.org/10.2139/ssrn.2697490.

Scherer, B. (2011). A note on the returns from minimum variance investing. *Journal of Empirical Finance*, *18*(4), 652-660.Available at: https://doi.org/10.1016/j.jempfin.2011.06.001.

Testoni, M. (2021). The market value spillovers of technological acquisitions: Evidence from patent-text analysis. *Strategic Management Journal*, 1-22.Available at: https://doi.org/10.1002/smj.3355.

**Appendix A.** Thomson Reuters company descriptions.

**Disney**: The Walt Disney Company, formerly TWDC Holdco 613 Corp, is a worldwide entertainment company. The Company operates in four business segments: Media Networks, Parks Experiences and Products, Studio Entertainment, and Direct-To-Consumer and International. The media networks segment includes cable and broadcast television networks, television production and distribution operations, domestic television stations, and radio networks and stations. The Company's Walt Disney Imagineering unit designs and develops new theme park concepts and attractions, as well as resort properties. The studio entertainment segment produces and acquires live-action and animated motion pictures, direct-to-video content, musical recordings and live stage plays. The Company also develops and publishes games, primarily for mobile platforms, books, magazines and comic books.

**Pfizer**: Pfizer Inc. (Pfizer) is a research-based global biopharmaceutical company. The Company is engaged in the discovery, development and manufacture of healthcare products. Its global portfolio includes medicines and vaccines. The Company manages its commercial operations through two business segments: Pfizer Innovative Health (IH) and Pfizer Essential Health (EH). IH focuses on developing and commercializing medicines and vaccines. IH therapeutic areas include internal medicine, vaccines, oncology, inflammation and immunology, rare diseases and consumer healthcare. EH includes legacy brands, branded generics, generic sterile injectable products, biosimilars and infusion systems. EH also includes a research and development (R&D) organization, as well as its contract manufacturing business. Its brands include Prevnar 13, Xeljanz, Eliquis, Lipitor, Celebrex, Pristiq and Viagra.

**JPMorgan Chase**: JPMorgan Chase & Co. is a financial holding company. The Company is engaged in investment banking, financial services. It operates in four segments, as well as a Corporate segment. Its segments are Consumer & Community Banking, Corporate & Investment Bank, Commercial Banking and Asset Management. The Consumer & Community Banking segment offers services to consumers and businesses through bank branches, automatic teller machines (ATMs), online, mobile and telephone banking. The Corporate & Investment Bank segment, comprising Banking and Markets & Investor Services, offers investment banking, market-making, prime brokerage, and treasury and securities products and services to corporations, investors, financial institutions, and government and municipal entities. The Commercial Banking segment provides financial solutions, including lending, treasury services, investment banking and asset management. The Asset Management segment comprises investment and wealth management.

**Bank of America**: Bank of America Corporation is a bank holding company and a financial holding company. The Company is a financial institution, serving individual consumers and others with a range of banking, investing, asset management and other financial and risk management products and services. The Company, through its banking and various non-bank subsidiaries, throughout the United States and in international markets, provides a range of banking and non-bank financial services and products through four business segments: Consumer Banking, which comprises Deposits and Consumer Lending; Global Wealth & Investment Management, which consists of two primary businesses: Merrill Lynch Global Wealth Management and U.S. Trust, Bank of America Private Wealth Management; Global Banking, which provides a range of lending-related products and services; Global Markets, which offers sales and trading services, and All Other, which consists of equity investments, residual expense allocations and other.