UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

CEFIN
Centro Studi Banca e Finanza

ISSN 2282-8168

CEFIN Working Papers No 105

Textual analysis in stock picking: additional evidence on (dis)similarity

by Riccardo Ferretti and Andrea Sciandra

August 2026

Dipartimento di Economia Marco Biagi – Università di Modena e Reggio Emilia
Via Berengario, 51- 41121 Modena, Italia
www.economia.unimore.it

Textual analysis in stock picking: additional evidence on (dis)similarity

Riccardo Ferretti[a,c,*]

Department of Communication and Economics, University of Modena and Reggio Emilia & Cefin - Centro Studi Banca e Finanza

Andrea Sciandra[b,d]

Department of Philosophy, Sociology, Education and Applied Psychology & Department of Statistical Sciences, University of Padova

[a] Department of Communication and Economics, University of Modena and Reggio Emilia, Viale Antonio Allegri 9, 42121 Reggio Emilia (RE), Italy. email: riccardo.ferretti@unimore.it

[b] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Via Cesarotti, 10/12, 35123 Padova (PD), Italy. email: andrea.sciandra@unipd.it

[c] CEFIN – Centro Studi di Banca e Finanza, Department of Economics Marco Biagi, University of Modena and Reggio Emilia, Viale Jacopo Berengario 51, 41121 Modena (MO), Italy.

[d] Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova (PD), Italy.

* Corresponding author

Declarations of interest: All authors declare that they have no conflicts of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Credit Author Statement:

Conceptualization: Ferretti R.

Methodology: Ferretti R., Sciandra A.

Software (computer code): Sciandra A.

Formal analysis: Ferretti R., Sciandra A.

Data Curation: Ferretti R.

Writing - Original Draft: Ferretti R., Sciandra A.

Writing - Review & Editing: Ferretti R., Sciandra A.

Textual analysis in stock picking: additional evidence on (dis)similarity

Abstract

This study investigates the predictive capability of textual similarity within the context of financial news, drawing upon empirical evidence from prior research. Two hypotheses are formulated and subsequently validated through analysis: firstly, that the similarity among financial newspaper articles concerning the same company can forecast future stock performance, and secondly, that a negative relationship exists between textual distance (TD) and future returns. The analysis demonstrates that employing textual similarity as a selection criterion can result in effective stock selection. This observation holds true across various similarity measures and remains robust even in out-of-sample tests. Additionally, the study emphasizes the significance of specific similarity measures, specifically Minmax and Jaccard, in predicting abnormal stock returns. Finally, we assess the ability of TD to construct stock portfolios capable of generating excess performance, finding that the low TD portfolio consistently exhibits statistically significant excess performance compared to the high TD portfolio.

Keywords: textual analysis; similarity measures; stock picking; return forecasting; investment decisions

JEL: G11; G17

1. Introduction and research hypotheses

Stock price forecasting has always been a topic of great interest to academics and practitioners, both as market timing (beta-predictability) and as stock picking (alpha-predictability)[1]. A strand of the literature employs textual analysis as a forecasting tool. Our study aligns with this research stream, focusing on medium to long-term forecasting, specifically a twelve-month horizon. This timeframe is typical for target prices formulated by financial analysts to inform stock picking.

The utilization of text analysis methodologies has seen a surge in popularity owing to the efficacy demonstrated by its applications over the past decade. These applications span from analyzing the content of social media messages to the latest advancements, which encompass Large Language Models (LLMs) and their derivative applications such as ChatGPT. As for intertextual distances, which constitute the focal point of this study, they are commonly employed for author recognition in the Humanities and in forensic applications (Eder et al. 2016), primarily from a

[1] For a review of the subject, consult, among others, Rapach and Zhou (2013) and Reschenhofer et al. (2020). Sonkavde et al. (2023) analyze the application of machine learning and deep learning models for stock price prediction. Recent papers on stock return forecasting include McMillan (2021), Stalla-Bourdillon (2022), and Farmer et al. (2023).

classification standpoint. In contrast, our study employs intertextual distances in the context of stock returns, presenting a distinct perspective on their usage.

Textual analysis has now carved out a large space in the social sciences, including in the financial and accounting fields. This is evidenced by the numerous reviews on the subject (Bochkay 2023; Das 2014; Gentzkow et al. 2019; Huang et al. 2020; Kearney and Liu 2014; Lewis and Young 2019; Li 2011; Loughran and McDonald 2016; 2020; Muslu et al. 2015; Xing et al. 2018). Regarding specifically the effectiveness of textual analysis to predict stock returns, Silva de Carvalho et al. (2023, 140) conclude their review with these words: «The results show evidence that textual sentiment can predict the return on stocks». Two relevant aspects emerge from this review: on the one hand, the studies cover a wide range of information sources (financial news, corporate disclosures, social media, and other documents); on the other hand, the analysis of the texts mainly concerned the measurement of sentiment or tone.

A part of the literature has dealt with changes to the language and construction of financial reports (similarity). Our paper also relies on similarity measures (which we refer to as Text Distance), but in the context of financial news, texts much less structured than earnings reports and with irregular frequency. Therefore, our main research question is as follows: does the predictive power of similarity apply only to corporate disclosure or also to financial news? Based on the arguments explained in the next section, we hypothesize that:

H1. The Text Distance of newspaper articles related to the same company can predict the upcoming performance of the underlying stocks.

H2. There is a negative relationship between Text Distance and future returns.

H3. The prediction effectiveness is higher for longer horizons.

Of course, verifying these hypotheses encounters an obstacle related to the nature of the textual source: when comparing two newspaper articles concerning a given firm published at different times, the degree of similarity is influenced not only by a change in language but also by the fact that the subject matter of the two articles may be completely different. In the cases analyzed by Cohen et al. (2020), all the texts examined shared the same subject: the earnings report. To overcome this issue, our sample of texts consists of articles belonging to the same column, thus sharing an identical format and type of information.

The results of our analyses confirm the three hypotheses. The subdivision of articles based on Text Distance (TD) allows for effective stock picking. In univariate analysis, the buy-and-hold abnormal return (BHAR) of the low TD stock portfolio exceeds that of the high TD stock portfolio, both in-sample and out-of-sample approaches, especially over the annual holding period. Multivariate analysis highlights a negative and statistically significant relationship between article TD and one-year BHAR, considering both characteristics of the company (such as size, beta, price-to-book value) and other textual characteristics, such as sentiment and uncertainty. Finally, time-series analysis reveals that the low TD portfolio generates, in the out-of-sample approach, a significant monthly alpha ranging from 43 to 78 basis points, while the high TD portfolio's alpha is statistically equal to zero (with the in-sample approach, the low TD portfolio's alpha varies from 38 to 99 basis points). Among the various distance measures used, Minmax and Jaccard have proven to be the most performant.

These findings contribute to the literature regarding the relationship between textual analysis and stock selection in two significant ways. Firstly, they validate the significance of text similarity in forecasting the future returns of individual stocks, even when dealing with poorly structured textual sources such as financial news, an area that has previously remained largely unexplored. Secondly, they provide novel comparative evidence regarding the efficacy of various similarity measures.

The structure of the paper is as follows. Section 2 outlines the conceptual and methodological framework. Section 3 describes the sample construction and data. Section 4 provides details on how text distance is measured. Section 5 presents the results of the univariate analysis, i.e. the BHARs of portfolios with text distances above or below the median. Section 6 reports the multivariate analysis examining the relationship between BHARs and text distance. Section 7 focuses on the measurement of alpha for portfolios with high and low text distances, while Section 8 serves as the conclusion.

2. Literature review and hypothesis development

The use of textual analysis in studying stock returns has been a prominent area of financial research for over a decade. Much of the focus has been on the sentiment or tone conveyed by the text. Notable studies in this area include Chen et al. (2014) and Sanford (2020) on social media; Feldman et al. (2010) on annual reports; Price et al. (2012) on conference calls; Tetlock (2007; 2011), Garcia (2013), Calomiris and Mamaysky (2019), Ke et al. (2019), and Liu and Han (2020) on newspaper articles. However, there are also studies that have considered the short-term market reaction to other textual aspects such as the levels of uncertainty in the texts of forms filed before an Initial Public Offering (Loughran and McDonald 2013), the frequency, and unusualness (entropy) of word flow in news articles (Calomiris and Mamaysky 2019), the managerial style in earnings conference calls, either in terms of vagueness (Dzieliński et al. 2017), or in balancing objective facts and subjective opinions (Kogan and Meursault 2021). More recently, textual analysis has been employed to develop a narrative-based factor pricing model using newspaper texts (Bybee et al. 2023).

Another branch of research has explored the predictive power of similarities between temporally consecutive texts, which is the focus of our study.[2] Cohen et al. (2020, 1413) found a significant predictive ability of textual (dis)similarity among earnings reports (form 10-K and 10-Q): «a portfolio that shorts “changers” and buys “nonchangers” in annual and quarterly financial reports earns 30 to 50 basis points per month over the following year», approximately up to 6% per year. Adosoglou et al. (2021) found that a portfolio of ‘‘non-changers" stocks (firms that do not change their 10-Ks in a semantically important way from the previous year) earns up to 10% in yearly alpha. The relationship between changes in balance sheet narratives and stock market reactions was previously highlighted by Brown and Tucker (2011), though with a different finding: "textual changes" were associated with higher returns. It is important to note that this earlier study, unlike more recent evidence, exclusively analyzed annual reports (specifically the Management Discussion and Analysis, or MD&A, section) and captured immediate market reactions (three-day stock returns) rather than longer-term outcomes. Further analysis by Brown and Tucker led to two noteworthy conclusions: (a) the MD&A section provides valuable information for forecasting cash flows over future periods but is less effective for predicting earnings in the immediate term; and (b) short-term price reactions to textual changes have diminished over time. This evidence suggests that changes in the textual content of financial statements take longer to influence investors’ revisions of expectations regarding business fundamentals.

Why should similarity possess the capability to forecast future equity returns? What should be the expected sign of this relationship? According to Cohen et al. (2020), while the immediate market response to quarterly and annual report releases may be less pronounced in contemporary times compared to the past, the informational value of these reports has not waned. The increasing complexity and length of such documents have fostered a pervasive form of investor inattention, impeding the prompt assimilation of information into stock prices, thereby rendering their returns foreseeable in the medium to long term. Specifically, investors often overlook subtle linguistic changes that foreshadow negative equity returns and adverse operational performance (to mitigate the risk of being sued for omitting negative events, language changes tend to introduce elements of negativity). The comparative analysis of textual data is less straightforward than that of numerical data, contributing to investors' reluctance to extract information from the former as readily as from the latter. Unlike numerical data presented alongside previous periods in financial statement tables, textual components lack this comparative framework, thus exacerbating investors' inertia in

[2] Similarity is also measured on contemporary texts. For example, Dyer et al. (2024) use similarity in disclosure between pairs of companies. In their case, similarity not only predicts but influences co-movements in the returns of the two companies and is useful for more accurate estimates of forward-looking market betas.

processing such information. Similar arguments are also advanced in Adosoglou et al. (2021). The notion that changes in the textual content of financial statements, specifically the MD&A section, convey new and hard-to-grasp information was first highlighted by Brown and Tucker (2011). From a general perspective, Enke et al. (2024) show that information-processing constraints lead to an insufficient sensitivity of people's decisions to economic fundamentals—a phenomenon they term Behavioral Attenuation.

Unlike the textual components of financial reports, press articles are neither as complex nor as lengthy. This raises doubts about whether the arguments made by Cohen et al. (2020) and Adosoglou et al. (2021) regarding financial reports can also apply to financial news. In the newspaper's context, Tetlock (2011) argues that textual similarity in financial news—defined as similarity to the preceding ten stories about the same firm—indicates the staleness of a news story (see also Ke et al. 2019). On news release days, the market reacts to both new and stale information; however, the response to second-hand news is smaller and dissipates within the following week. Notably, the reversal effect becomes stronger with increasing similarity. This evidence demonstrates that similarity predicts return reversals in the very short term, while implicitly suggesting that stale information (characterized by high similarity) lacks predictive power over the long term. In contrast, other studies have found that the textual flow of news articles—focusing on aspects beyond similarity—is more effective for making long-term predictions, such as those spanning a year, than for shorter-term predictions like monthly forecasts (Calomiris and Mamaysky 2019; Glasserman and Mamaysky 2019).

The articles in our sample consist solely of stale information. They simply describe the business profile and performance of specific companies and are not driven by corporate press releases (see the next section for details). From this perspective, we should not expect to observe any relationship between similarity and future stock returns within our sample. However, there are reasons to consider this possibility. First, while our articles are certainly less complex than annual or quarterly reports, they are still significantly longer than typical news articles. Second, the comparison is made with a prior article about the same company, published in the same column, usually about a year earlier. Both factors make it challenging to immediately compare the texts. From an investor's perspective, it may take time for the informational content of the similarity to be fully reflected in market prices; quoting the words of Cohen et al. (2020, 1372): «investors uncover the implications of the news contained in document changes only gradually over time, but eventually the news is fully impounded into stock prices and firm operations».

But what could this informational content be? To address this question, we must adopt the perspective of journalists. Journalists, being in regular contact with managers, analysts, and institutional investors, are uniquely positioned to detect even the earliest signs of shifts in market perception regarding a company's prospects. These subtle changes, which are not yet fully priced into the current stock value, may eventually influence market dynamics. Changes in language and structure in an article, relative to its predecessor, translate into lower similarity. Acting on similarity rather than providing explicit buy or sell recommendations can be seen as a strategy to mitigate the risk of violating insider trading regulations. However, it remains to be determined whether the signals conveyed through similarity are positive or negative—that is, whether low similarity predicts negative future returns, as suggested by Cohen et al. (2020) and Adosoglou et al. (2021), or positive ones. Since modifying language and structure requires additional effort on the journalist's part, it is important to consider whether this effort is more likely to be undertaken in response to positive or negative perceptions. Two well-documented cognitive biases provide insight here. First, the *status quo bias*—the tendency to «do nothing or maintain one's current or previous decision» (Samuelson and Zeckhauser 1988, p. 7)—suggests a general preference for minimal change. Second, the *negativity bias*—the principle that «bad is stronger than good» (Baumeister et al. 2001, p. 323)—implies that negative developments are more likely to prompt action. These biases lead to the hypothesis that negative perceptions are more likely to induce variations in language and structure. Consequently, a positive relationship between similarity and future returns (or a negative relationship with our measure of Text Distance) is expected.

Based on the above-mentioned arguments, we hypothesize that lower similarity between an article and its predecessor is associated with negative future returns. This relationship arises because journalists are more likely to exert the effort to modify language and structure when they perceive negative developments, driven by the combined effects of status quo bias and negativity bias. Thus, greater textual differences (lower similarity or higher Text Distance) signal shifts in market perceptions that may foreshadow adverse outcomes for the company's stock performance. More precisely:

H1. The Text Distance of newspaper articles related to the same company can predict the upcoming performance of the underlying stocks.

H2. There is a negative relationship between Text Distance and future returns.

H3. The prediction effectiveness is higher for longer horizons.

In this perspective, high text similarity represents "good news" and vice versa. Therefore, a stock-picking strategy based on textual similarity involves long positions on stocks of companies with high similarity (low Text Distance) and short positions on stocks of companies with low similarity (high Text Distance).

Methodologically, Cohen et al. (2020) compute similarity using measures derived from studies in linguistics, textual similarity, and NLP, such as cosine similarity, Jaccard similarity, minimum edit distance, and simple similarity. Conversely, Adosoglou et al. (2021) employ neural network embedding techniques (Word2Vec and Doc2Vec algorithms) to capture changes in both syntax and semantics, to which they then apply the cosine similarity measure, the Euclidean distance, the Radius of Gyration, the Jaccard similarity; their «model focuses more on the changes in the topic covered and writing style in the 10-Ks by representing arbitrarily the entire documents with vectors in a fixed-dimensional semantic space». The use of large language models (LLMs) that allow representations of words dynamically informed by the words around them (Devlin et al. 2019) could improve this approach. However, currently, there are no pre-trained LLMs on financial texts in the Italian language. Therefore, we opted for an approach similar to that of Cohen et al. (2020) but using a wider range of distance measures, following the stylometric literature (Eder et al. 2016).

In the following sections, we will examine the predictive capacity of TD and the existence of an inverse relationship between TD and stock returns using three different methods: analysis of Buy and Hold Abnormal Return (BHAR) of high and low TD portfolios (event-time returns); cross-section regressions of individual stock BHAR on our TD measures and other control variables (cross-section regression); analysis of Calendar-Time Returns of high and low TD portfolios (time-series regression). These types of analyses, with some differences, are present in Cohen et al. (2020) and, limited to the first and the third, in Adosoglou et al. (2021).

3. Sample and data

In order to test our research hypotheses, we identified all the 809 articles belonging to the 'Letter to investor' (*Lettera all'investitore*) column published in the Sunday edition of the main Italian financial newspaper, *Il Sole 24 Ore*, from January 2005 to June 2022. The ‘Letter’ offers second-hand information about a specific listed company (in few cases two or more companies), such as balance sheet and income statement data, managers' outlook, stock's past performance, and occasionally, analysts' recommendations. After excluding 29 articles[3], we were left with 780 records, of which 179 were unique articles (i.e. related to companies mentioned only once) or first articles (i.e. articles not preceded by other articles related to the same company) for which it was obviously not

[3] We excluded articles covering companies listed on foreign exchanges (8) or with incomplete data (11). We also set aside articles mentioning more than one company (6) or companies involved in special situations such as M&As or insolvency proceedings (3) and articles whose file was missing in the newspaper's electronic database (1). Given that excluded articles account for only 3.6%, we believe their impact on results is negligible.

possible to calculate the distance between one text and another. From these 780 articles, 601 Text Distances were then derived relating to 135 listed companies.

It is evident that our sample is not random, posing a risk of sample selection bias. In essence, our findings may stem from undisclosed criteria used by article authors in selecting companies for analysis, rather than from their degree of Text Distance.

The repeated presence of a relatively small number of companies, on the one hand, and the regularity of the time interval between one article and another of the same company, on the other, suggest the willingness of journalists to cover already known companies and the limited influence of contingent factors such as market trends or attention to a certain sector or a particular company. Initial selections of companies for coverage likely weighed these factors more heavily than subsequent articles. Since our focus on Text Distance pertains to articles after the initial coverage, we anticipate minimal influence from these unobservable factors, mitigating the risk of endogeneity in our results. Recognizing that our reliance on articles solely from Il Sole 24 Ore restricts the generalizability of our findings, we have not found alternative columns with characteristics conducive to identifying Text Distance effects over an extended period. However, the opportunity to assess the efficacy of Text Distance in stock picking across a prolonged timeframe, encompassing alternating bear and bull markets, enhances the robustness of our results. Additionally, it is not uncommon in the literature to find studies analyzing single columns such as 'Herd on the Street' (Lloyd-Davis and Canes 1978; Syed et al. 1989; Pound and Zeckhouser 1990; Liu et al. 1990; 1992; Beneish 1991) or 'Dartboard' (Barber and Loeffer 1993; Albert and Smaby 1996; Pettengil and Clark 2001), all of which are published in the Wall Street Journal.

The average time interval between two articles covering the same company was 533 calendar days (392 days the median value), with a minimum of 119 days and a maximum of 3,794 days. In the almost eighteen years taken into consideration, three journalists have alternated, the first signed all the articles published up to February 27, 2011, the second almost all the articles from March 6, 2011, to February 5, 2012, the third most of the articles from February 12, 2012 to January 6, 2013 and all the subsequent ones. From March 2010 to February 2011 the column was discontinued. Of the 601 distances, 79 concern articles written by different journalists, 74 of which concentrated in the years 2011-2014. Figure 1 illustrates the temporal distribution of the 601 articles for which textual distance could be calculated. The limited number of articles in 2005 and 2010 stems from distinct reasons: in 2005, all but two were first articles; in 2010, the column was not published for ten months.

<Figure 2 about here>

For each article, we collected the following data:

(1) Name of the analyzed company.

(2) Size. Company size in terms of capitalization (natural logarithm), which is the company's market value as of one week before the publication of the article.

(3) PBV. Price-to-book value of the analyzed company as of one week before the publication of the article.

(4) Days. Number of calendar days elapsed with respect to the previous article concerning the same company.

(5) New. This dummy variable signals a different columnist with respect to the previous article concerning the same company (1 if a change in the article's author is observed).

(6) Recommendation. Presence in the article of one or more analysts' recommendations (a dummy variable equal to 1 if recommendations are reported).[4]

(7) TD. Text Distance (see section 4 for details), the higher the distance, the lower the similarity between the content of two consecutive article covering the same listed company. TD is our key variable for portfolio selection. The link between TD and portfolio performance is analyzed following three different approaches: event-time returns (section 5), cross-section regression (section 6), and time-series regression (section 7).

(8) TS. Text sentiment (see section 4 for details).

(9) TU. Text uncertainty (see section 4 for details).

To compute the following variables, data was obtained from Eikon:

(10) Monthly returns of the company's stock and the stock market index (FTSE Italia All-Share Index) from month +1 to month +12 relative to the publication month[5]. From these returns, Buy and Hold Abnormal Returns (BHAR) are obtained for each stock/article over four different time horizons: 3, 6, 9, and 12 months[6]. BHARs are our dependent variables in the univariate (section 5) and multivariate (section 6) analysis. The number of BHARs declines in the nine- and twelve-month horizons (from 601 to 583 and 546, respectively) since we removed all the return duplications that occur when the time between two articles covering the same company is less than twelve months. Monthly returns of the company's stock and the stock market index are also the basis for our time-series analysis (section 7).

[4] From 2005 to 2014, most of the columns featured analysts' suggestions (buy, hold, or sell). Starting from 2015, these explicit recommendations ceased to be included.

[5] We follow Cohen et al. (2020) in choosing to calculate returns from month +1 instead of the date of publication of the article.

[6] $BHAR_{i,N} = (\prod_{m=1}^{N}(1 + AR_{i,m})) - 1$ where $BHAR_{i,N}$ is the Buy and Hold Abnormal Return of the company's stock "*i*" over the *N*-month period and $AR_{i,m}$ is the return of the company's stock "*i*" in month "*m*" net of the market return in the same month.

(11) Beta. Market Model Company's Beta, a measure of the stock's sensitivity to market movements, estimated on the daily returns of the company's stock and the stock market index (FTSE Italia All-Share Index) from day -16 to day -215 (200 working days) relative to the publication day.

(12) PP. Market-adjusted company's past performance, calculated as the sum of daily stock returns from day -215 to day -16 before the publication day, adjusted for the market return during the same period.

(13) AR(0). Abnormal Return at t=0. Company's stock return in the first working day after the publication, adjusted for the market return in the same day.

Table 1 presents the descriptive statistics of the dependent and independent variables used in the regression analysis. The variable "PP" (12) is included in the analysis to account for the potential influence of non-random data collection. "Size" (2), "PBV" (3), and "Beta" (11) are used to capture the structural characteristics of the company that may impact investors' behavior. AR(0), the short-term market reaction to the column (13), is included since it could be a predictor of the long-term performance. Days (4), New (5), Recommendation (6), Text Sentiment (8), and Text Uncertainty (9) control for other characteristics of the column beside the Text Distance. For comparison's sake, the control variables in Cohen et al. (2020) are: standardized unexpected earnings surprise (SUE); size; log(1/PBV); previous month's return; cumulative stock return from month t − 12 to month t −2; sentiment of change; uncertainty of change; log(Document Size); quarter-on-quarter change in log(Document Size); change CEO/CFO.

<Table 1 about here>

4. Text analysis

To analyze the column texts, we first calculated intertextual distances. We used dissimilarity measures to compare the text of the article related to a given stock with the previous stock's text (when available). The procedure for this operation involves pre-processing the article texts. This includes reducing all words to lower case, removing punctuation, numbers, and Italian stop-words. The document-term matrix (DTM) was then produced from these steps. We decided not to proceed with text lemmatization to emphasize the importance of verb tenses. In this type of article, for example, it is crucial to write 'it succeeded' instead of 'it will succeed' to convey the intended meaning to the reader.

To calculate stylometric distance metrics, we utilized the R package `stylo` (Eder et al. 2017) and integrated it with ad hoc code to compute additional metrics not found in this library. Initially, mainly following Eder et al. (2017), we selected these distance metrics between the frequency patterns of individual stocks in our corpus:

- Argamon' Delta (Argamon 2008)
- Burrows's Delta (Burrows 2002)
- Cosine (Evert et al. 2017)
- Eder's Delta (Eder 2015)
- Entropy (Joula and Baayen 2005)
- Jaccard (Jaccard 1901)
- Minmax (Koppel and Winter 2014; Kestemont et al. 2016)
- Simple Delta (Eder 2015)
- Wurzburg (Jannidis et al. 2015).

These distances provide a comprehensive basis for analyzing differences between texts, as they account for almost all of those used in stylometry (Eder et al. 2016). Since all of these distances were found to be significantly correlated with each other, we selected the two distances with the highest significant correlation coefficient with Buy and Hold Abnormal Returns at 12 months (BHAR(12)), namely Jaccard and Minmax (r = -0.175*** and -0.176***, respectively), along with the distance that had the lowest significant correlation coefficient with BHAR(12), which was Burrow's Delta (r = -0.099*). Therefore, our choice is based on a clear criterion, namely maximizing the correlation with the dependent variable, represented in this case by BHAR(12). Moreover, as previously indicated, these distances exhibit highly significant and strong correlations between them, with a minimum of 0.469 ($p < 0.001$) between the Eder distance and the Delta distance and a maximum of 0.995 ($p < 0.001$) between the Entropy distance and Simple distance. Although the Jaccard and Minmax distances also show a very high correlation (0.979, $p < 0.001$), we decided to utilize both in subsequent analyses because their correlation with BHAR(12) was the highest and quite similar. Even if we expected better results for the Minmax distance, we wanted to test both distances to observe their association with other control variables that might change the intensity and significance of their relationship with BHAR(12) (see Section 6 in particular). The selection of the Delta distance represents a form of control analysis, as it exhibits the least significant correlation with BHAR(12). However, we sought to observe its behavior in association with other control variables with respect to the relationship with BHAR(12) (see Section 6 in particular).

We calculated the selected distances as follows. Let X and Y be two document vectors, with n features. Let x_i and y_i represent the value of the i-th feature in both documents respectively (e.g., the relative frequencies), then:

$$Delta\_distance(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\mu_i}{\sigma_i}-\frac{y_i-\mu_i}{\sigma_i}\right|, \tag{1}$$

where μ_i is the mean frequency of the *i*-th feature in the corpus, and σ_i is the standard deviation of frequencies of the *i*-th feature.

$$Jaccard\_distance(X,Y) = 1 - J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

$$Minmax\_distance(X,Y) = 1 - Minmax = 1 - \frac{\sum_{1=1}^{n}\min(x_i,y_i)}{\sum_{1=1}^{n}\max(x_i,y_i)} \tag{3}$$

Delta_distance (Delta) relies on *z*-scores (i.e. normalized word frequencies) and it is dependent on the number of texts analyzed. Therefore, there must be a balance between these texts (Eder et al. 2017), as in our case, since the column's articles showed an average length of about 1500 words and the column maintained almost the same layout over the analyzed period.

Jaccard measure ($J(X,Y)$) was proposed in 1901 (Jaccard 1901) as a tool to detect co-localization of alpine flora (Costa 2021) and it is defined as the intersection of two documents divided by the union of those two documents. Since the Jaccard measure ranges from 0 to 1, then also *Jaccard_distance* (Jaccard) ranges from 0 to 1 and it measures the dissimilarity between two document vectors.

Minmax measure was originally introduced in geo-botanics by Ružička (1958) and recently employed in authorship attribution (Kestemont et al. 2016). *Minmax_distance* (Minmax) is very similar to the Jaccard distance (Costa 2021; Kestemont et al. 2016), so we expect similar results when using these two distances to estimate BHAR.

Overall, the Jaccard distance is well known and used in various fields, while it is less common to find applications on intertextual distances using Minmax. Delta distance, on the other hand, is very successful in stylometry, also from a theoretical point of view (Eder et al. 2016).

From a technical viewpoint, the procedure for calculating distances involved the following steps:

- The reduction of all words to lowercase.
- The elimination of punctuation, numbers, and symbols.
- The removal of Italian stop words (in order to use only meaningful words).

We also decided not to remove low-frequency words, as they might represent possible markers of the presence of new concepts or judgments with respect to the analyzed stocks.

Regarding sentiment analysis, in previous works we used a general lexicon, pointing out the need for resources in Italian comparable to Loughran and McDonald's (2011) financial lexicon. This lexicon includes lists of positive and negative terms, as well as other potentially useful word lists, such as that related to uncertainty. Therefore, we chose to use the 'eTranslation' online machine translation service provided by the European Commission (https://commission.europa.eu/resources-partners/etranslation_en) to automatically translate the Loughran and McDonald's lexicon. We then conducted a qualitative review of the results. It is important to note that in this analysis, the term 'sentiment' refers to the tone of the article and should not be confused with the concept of investor sentiment. According to Baker and Wurgler (2006), investor sentiment refers to the belief about future cash flows and investment risks that is not justified by the facts at hand, or as the propensity to speculate, or as the overall optimism or pessimism about a given asset. Our approach differs from the traditional method of constructing an aggregated investor sentiment index through principal component analysis of underlying economic variables such as retail investor trades, trading volume, option implied volatility, and first day returns on initial public offerings (IPOs). In this study, we focused on a text-based document-level sentiment indicator using an unsupervised lexicon-based approach. Thus, in a bag-of-words framework, sentiment analysis was conducted by considering each column as a combination of its component words, producing a polarity score for each article. The Loughran and McDonald's list of words related to uncertainty was also exploited in the same way, providing an uncertainty score for each article.

<Figure 2 about here>

Figure 2 illustrates the temporal dynamics of the three distance measurements mentioned above. Two notable circumstances emerge: (i) a significant spike in the years 2011 and 2012; (ii) average values from 2005 to 2010 surpass those from 2015 to 2022. The first circumstance relates to the turnover in column editors, resulting in a notable increase in dissimilarity: all articles from 2011 and 2012 were penned by a different journalist compared to the preceding articles covering the same company. This percentage decreases to 42% and 15% in 2013 and 2014, respectively. The second circumstance hints at a potential disparity in style between the two primary editors: one responsible for all articles published up to February 27, 2011, and the other for those since January 6, 2013. Furthermore, considering the discontinuation of analysts' suggestions (buy, hold, or sell) in January 2015, it is pertinent to examine not only the 2005-2022 timeframe but also the narrower and more homogeneous range of 2015-2022 to mitigate potential distortions in the distance measures arising from these circumstances.

5. Event-time returns from TD strategies

In this section, we examine the abnormal cumulative returns of stock-picking strategies, both in sample and out-of-sample, utilizing TD as a selection tool. For each of the 601 articles under consideration, the following metrics were computed: (a) the textual distance from the preceding article pertaining to the same company (TD_i); (b) the Buy and Hold Abnormal Return ($BHAR_i$) of the corresponding stock over 3 to 12-month horizons from the month of publication. After determining the median of the 601 distances, articles with TD_i below the median were incorporated into the Low Text Distance (LTD) portfolio, while those with TD_i above were included in the High Text Distance (HTD) portfolio. Subsequently, for each of the two portfolios, four $BHAR_p$ were calculated —one for each time horizon— as a simple average of the $BHAR_i$ values of the individual stocks comprising the reference portfolio, in symbols:

$BHAR(T)_p = \frac{\sum_{i=1}^{Np} BHAR(T)_i}{N_p}$, where

$BHAR(T)_p$ = Average Buy and Hold Abnormal Return of portfolio *P* (with *P*=LTD or HTD) over horizon *T* (with T = 3, 6, 9, 12);

$BHAR(T)_i$ = Buy and Hold Abnormal Return of stock *i*, over horizon *T*, included in portfolio *P*;

N_p = number of stocks in portfolio *P*.

Table 2 presents the results of the in-sample analysis using three distinct TD measures: Minmax, Jaccard, and Delta. The articles span the entire period from January 2005 to June 2022. The table includes the total number of observations (articles), the number of observations and BHARs for each portfolio, the calculation of the differential returns ($\Delta = BHAR_{LTD} - BHAR_{HTD}$), and the corresponding significant tests (t-test; Mann-Whitney U test; Kolmogorov-Smirnov test).

<Table 2 about here>

Notably, the LTD portfolio consistently outperforms the HTD portfolio, particularly evident from the 6-month horizon onward for both Minmax and Jaccard measures. The magnitude and significance peak at the 12-month horizon, showcasing a substantial differential return of 9.50% and 9.11%, contingent on the TD metric used (Minmax or Jaccard). Predictive efficacy is also found to be higher in the long-term in the work of Brown and Tucker (2011) and Calomiris and Mamaysky (2019), who respectively classified the context and content of news articles to predict risk and return in equity markets in 51 developed and emerging economies, and showed that economic and statistical significance is high and higher for one-year forecasts than for monthly forecasts.

The out-of-sample analysis was conducted using two methods to calculate the median TD. In the first method (Table 3), distances related to articles published in the years 2005-06 were utilized to compute the median distance for categorizing the initial article of 2007[7]. Subsequently, the TD of each new article was compared with the median calculated across all TD values from January 2005 to the present moment (rolling basis median). In the second method (Table 4), distances from the 2005-06 period were employed to compute the median distance for classifying all articles from 2007. Starting from 2008, the TD of each new article was compared with the median of the preceding calendar year (yearly revised median). Both methods are ex-ante and out-of-sample, indicating that the classification of an article is determined solely based on information available at the time of portfolio attribution. The number of observations is lower compared to the in-sample approach because we lose the BHARs of articles from 2005 and 2006.

<Table 3 about here>

Tables 3 and 4 demonstrate that even in an ex-ante setting, the LTD portfolio outperforms the HTD portfolio in terms of abnormal returns, particularly in longer horizons and when utilizing the Minmax measure of TD. In our data, the differential return is due to the positive abnormal return of the LTD portfolio, while in Cohen, Malloy, and Nguyen (2020) it is due to the negative performance of the HTD portfolio. The yearly-revised method (Table 4) yields higher and more statistically significant differential returns compared to the rolling-basis method (Table 3). Specifically, the number of differential returns that are statistically significant in all three tests is 8, as opposed to 3 in the rolling-basis method. Additionally, the Δ BHAR(12)s are higher with the yearly-revised method: 9.28% vs. 6.86% with Minmax, 9.07% vs. 5.28% with Jaccard, and 8.41% vs. 6.24% with Delta.

<Table 4 about here>

<Table 5 about here>

To assess the robustness of our results, we performed the ex-ante exercise over a limited period, commencing from 2015. This solution aims to mitigate potential distortions in the TD measures arising from changes in the journalist authoring the column, particularly observed in the 2011-2014 period (as discussed in sections 2 and 3). The latest data substantiates the existence of a performance differential in favor of the Low Text Distance (LTD) portfolio, albeit statistically significant solely within the 12-month horizon, employing both the 'rolling' method (Table 5) and the 'yearly revised' method (Table 6). In contrast to preceding results, the Jaccard distance metric predominates over Minmax, and the 'rolling' method prevails over the 'yearly revised' method: with the rolling median, Δ BHAR(12) stands at 16.36% for Jaccard and 13.87% for Minmax; with the yearly revised median, Δ BHAR(12) is 12.33% for Jaccard and 11.83% for Minmax.

[7] The initial median is computed using data from the 2005-06 biennium, given that only two TDs correspond to articles published in 2005.

<Table 6 about here>

6. Cross-section regression: TD and stock performance

This section presents the regression model results for the Buy and Hold Abnormal Return ($BHAR_i$) of the corresponding stocks over a 12-month horizon from the month of publication. The focus is on testing the significance of the intertextual distance measures in these models. OLS regression and regression models with backward selection were used to analyze the contribution of distance measures in explaining BHARs. Backward selection is commonly used to simplify a regression model by reducing the set of predictors to only those that are significant, without excessively affecting performance. We chose a selection process based on minimizing the Akaike Information Criteria (AIC), which quantifies the amount of information lost during selection by means of log-likelihood[8].

Two tests were conducted: one using the complete dataset from 2005 and the other using only the stocks mentioned in the column from 2015. Initially, the OLS models included all variables presented in Section 3. Regarding the measures of intertextual distance, each model uses one of the three selected and described in Section 4 (Minmax, Jaccard, and Delta), along with their interactions with the dichotomous variable 'New,' which represents the journalist's change from the previous article related to the same stock.

<Table 7 about here>

Tables 7 and 8 show the results of these models for data from 2005 and 2015, respectively. The estimated coefficients, standard errors, and significance levels for each explanatory variable are presented. Additionally, the adjusted R-squared, *F*-statistics, and number of observations are provided for the overall models.

<Table 8 about here>

Table 7 indicates that both the Minmax and Jaccard distances are significant in both the OLS and backward models, while the Delta distance is never significant. The coefficients of the Minmax and Jaccard distances have negative signs, indicating that as the distance increases, the value of

[8] It is well known that biased estimates of standard errors of coefficient estimates can occur with this type of Backward Elimination (BE) (Smith 2018). However, it should be noted that: (i) the bias may be modest, as in our case, where the coefficients and standard errors of the OLS and BE regressions are very similar (Tables 7 and 8); (ii) although procedures such as BE may lead to biased estimates, they will have lower variance and fewer assumptions than other methods (Hastie et al 2009) (iii) we used BE to select the best subset of variables based on AIC criterion, which is equivalent to leave-one-out cross-validation.

BHAR(12) decreases, as expected. The interactions between distances and the presence of a new journalist never turn out to be significant.

We also performed some tests to support the identified causal mechanisms, in particular by exploring designs with an instrumental variable (IV). To this end, we chose to compute a classic measure of lexical diversity, the Type-Token Ratio (TTR), for each article since 2005. Then, we tested whether each distance considered was endogenous with respect to the BHARs at 12 months by fitting an instrumental-variable regression (Green, 1993) by two-stage least squares (2SLS). The results show that the *F*-test for weak instruments (first stage of the 2SLS estimation) was always significant ($p<0.001$ for Jaccard and Minmax distances; $p<0.01$ for Delta distance), so the TTR turns out to be a relevant IV for the distances. Furthermore, through the Wu-Hausman test, we compared the IV estimates with OLS estimates to find endogeneity. This test did not prove significant ($p>0.05$) for any distance, suggesting no evidence of endogeneity and confirming the choice to use OLS regressions.

All models in Table 7 are significant under *F*-statistics, and the backward model that includes Minmax distance has the highest adjusted R-squared value (5.7%). Our R-squared are comparable to those reported by Cohen et al. (2020; Tables IV and V).

Even for models that only include observations from 2015 onward, we find the same significant results as for the entire period (Table 8). The Minmax and Jaccard distances are always significant and have a negative sign, while the Delta distance is never significant. Although the number of observations is smaller, all models in Table 8 are significant, according to *F*-statistics. The backward regression, which includes Minmax distance, is again preferable due to the highest adjusted R-squared value (7.5%).

We also conducted the same analysis on time horizons shorter than 12 months (3, 6, and 9 months), but found little significance for the distances. Considering the entire observation period, the Minmax distance is significant at a 10% level, along with its interaction with the variable 'New' over a 6-month horizon, but only for the OLS model. Additionally, the interaction between Jaccard distance and 'New' is also significant (10%) on a 6-month horizon for the OLS model. On the other hand, the Delta distance is only significant in the interaction with the 'New' variable at a 3-month horizon, in both the OLS and backward models. In summary, our findings indicate that the Minmax and Jaccard distances are consistently significant in BHAR regressions over a 12-month horizon.

7. Time-series regressions of portfolio returns: looking for alpha

The objective of this section is to assess whether employing Text Distance (TD) allows for the construction of a stock portfolio capable of generating extra-performance in addition to the compensation for systematic risk (the so-called alpha coefficient). By applying a pricing model, we aim to verify the presence of a positive and significant intercept. To enhance the robustness of the results, we utilize two pricing models:

- the standard Market Model, where $R_p = \alpha_p + \beta_p \cdot R_m$ (A)
- the three-factor Market Model, where $R_p = \alpha_p + \beta_p \cdot R_m + B_{p2} \cdot VmG + B_{p3} \cdot SmB$ (B)

Here, R_p represents the monthly return of portfolio *P*, R_m is the monthly return of the market index (FTSE Italy All-Share Index); VmG (Value minus Growth) is the difference between the monthly returns of the Value stock index (MSCI Italy Value Price Index Local End of Day) and the Growth stock index (MSCI Italy Growth Price Index Local End of Day); SmB (Small minus Big) is the difference between the monthly returns of the Small Cap stock index (MSCI Italy Small Cap Price Index Local End of Day) and the Large Cap stock index (MSCI Italy Large Cap Price Index Local End of Day), α_p is the alpha of portfolio *P*; β_p, β_{p2}, and β_{p3} measure the sensitivity of portfolio *P* to the respective systematic risk factors. The VmG (value) and SmB (size) factors are based on the well-known three-factor model by Fama and French (1993).[9]

Two approaches are employed for portfolio composition. In the first approach, each month sees the creation of two equally-weighted portfolios: (i) the Low Text Distance (LTD) portfolio, comprising stocks discussed in the 'Letter to Investor' released that month, characterized by a TD lower than the current year's median; (ii) the High Text Distance (HTD) portfolio, encompassing stocks covered in the 'Letter to Investor' released that month, characterized by a TD higher than the current year's median. Each stock remains in the portfolio from the month immediately following the article's publication until the month it is mentioned again in another article from the column, with a maximum holding period of twelve months. This approach follows an in-sample (ex-post) perspective since the median TD value is not known at the time of classification but only at the end of the year.

In the second approach, portfolios are constructed each month as follows: (i) the LTD portfolio, including stocks discussed in articles released that month with a TD lower than the previous year's median; (ii) the HTD portfolio, comprising stocks discussed in articles released that month with a TD higher than the previous year's median. In line with the first approach, each stock remains in the portfolio from the month immediately following the article's publication until the month it is

[9] Cohen et al. (2020) apply both the 3-Factor Fama and French model (market, size, and value factors) and the 5-Factor model (market, size, value, momentum, and liquidity factors). Adosoglou et al. (2021) utilize the CAPM, the 3-Factor Fama and French model, and the 5-Factor model.

mentioned again in another article from the column, with a maximum holding period of twelve months. This approach follows an out-of-sample (ex-ante) perspective since the median TD value is known at the time of classification and entails a reduction in the temporal horizon under analysis, starting not from January 2005 but from January 2007 (the first median is calculated based on data from the 2005-06 biennium as only two TDs pertain to articles published in 2005). To make the results of the two approaches more comparable, the time series of returns for the in-sample approach also starts from January 2007. Among the different measures of TD, we exclusively utilize Minmax, identified as the best-performing option in the regression analysis.

In section 2, we remarked that the column has been curated by different journalists, with the alternation mainly occurring in the period 2011-2014. These discontinuities may have potentially distorted, likely on the upside, the Text Distance of the articles affected by the journalist change. For this reason, the excess returns of portfolios with low and high distance were also estimated over a restricted period (from 2015), during which distances were calculated exclusively based on articles by the same author. In the in-sample approach, the average number of shares in the LTD portfolio is 16.7 during the extended period and 18.9 during the short period, with corresponding values for the HTD portfolio being 17.3 and 19.1, respectively. In the out-of-sample approach, the average number of shares included in the LTD portfolio is 17.8 during the extended period and 19.6 during the narrow period, while the corresponding values for the HTD portfolio are 15.3 and 18.5.

<Table 9 about here>

Table 9 presents the results of in-sample estimates for the period January 2005-June 2022 (Panel A) and the period January 2015- June 2022 (Panel B); columns (4) to (6) refer to the standard Market Model, while columns (7) to (9) pertain to the three-factor Market Model. Columns (4) and (7) concern the portfolio composed of stocks with a TD below the median (LTD), columns (5) and (8) relate to the portfolio composed of stocks with a TD above the median (HTD), and columns (6) and (9) pertain to the portfolio with a long position on LTD stocks and a short position on HTD stocks (LTD-HTD).

Looking at Panel A, for both pricing models, while the LTD portfolio demonstrates the ability to generate statistically significant alpha, the same cannot be said for the HTD portfolio: 0.515% (column (4)) versus 0.130% (column (5)) according to the standard Market Model and 0.376% (column (7)) versus 0.011% (column (8)) according to the 3-factor Market Model. The alphas of the Long-Short portfolio (0.385% and 0.365%) are also statistically significant at the 10% level (columns (6) and (9)). This asymmetry in the performance of the two portfolios is also found in Liu and Han (2020) with reference to low- and high-negative tone portfolios: only the portfolio with low-negative tone has a significant return. In Cohen et al. (2020), the alphas of both portfolios are significant but with the opposite sign: positive for the LTD portfolio and negative for the HTD portfolio.

The outperformance of the LTD portfolio is even more pronounced in the data from January 2015 (Panel B): the alpha increases to 0.989% with the standard Market Model (column (4)) and 0.773% with the 3-factor Market Model (column (7)). The alpha of the HTD portfolio remains statistically indistinguishable from zero. The excess returns of the Long-Short portfolio increase to 0.835% (column (6)) and 0.786% (column (9)) in the first and second models, respectively. Compared to a passive portfolio with equal systematic risk, TD-guided stock picking has thus generated an economically significant annualized excess return: between 6.18% and 4.51% from 2005 and between 11.87% and 9.28% from 2015, depending on the pricing model.

Turning to the out-of-sample approach (Table 10), the effectiveness of TD as a selection tool is robustly confirmed. The LTD and HTD portfolios exhibit divergent behaviors irrespective of the pricing model and time span, solely the former manifests a positive and statistically significant excess performance. Over the extended period (Panel A), it registers 0.551% versus 0.178% (MM, columns (10) and (11)), and 0.428% versus 0.064% (3-factor MM, columns (13) and (14)). Over the restricted period (Panel B), the figures stand at 0.777% versus 0.211% (MM) and 0.554% versus 0.029% (3-factor MM). The Long-Short portfolio (LTD-HTD) exhibits an alpha ranging from 0.37% (Panel A, columns (12) and (15)) to 0.566%/0.525% (Panel B), with the latter being statistically significant.

<Table 10 about here>

Overall, the values we find are on the same order of magnitude as the alphas reported by Cohen et al. (2020) and Adosoglou et al. (2021) concerning the similarity between earnings reports. Moreover, they are equal to or greater than the alphas generated by stock-picking strategies driven by sentiment in social media (Chen et al. 2014) and newspapers (Liu and Han 2020).

8. Concluding remarks

This study offers comprehensive evidence in support of utilizing textual similarity (Text Distance – TD), particularly metrics like Minmax and Jaccard, as robust predictors of future stock performance. It adds valuable insights to the literature by affirming the significance of textual analysis in forecasting stock returns, even in the domain of financial news, and by presenting comparative evidence on the effectiveness of various similarity measures.

This research delves into exploring how textual similarity within financial news articles can function as a predictive tool for future stock performance. Building upon existing empirical research, the study posits two hypotheses: firstly, that the level of similarity among newspaper articles related to the same company can predict the forthcoming performance of associated stocks; and secondly, that there is a negative correlation between textual distance and future returns. Through analyses, the study confirms both hypotheses. It illustrates that segmenting articles based on their textual similarity facilitates effective stock selection. Particularly over extended time horizons, we observed the significance of Minmax and Jaccard TD in predicting abnormal stock returns. In contrast to Cohen, Malloy, and Nguyen (2020), where all similarity measures exhibit significant coefficients in multiple regression, our study finds Delta TD to be nonsignificant, as expected, given its poor correlation with BHARs.

We looked for alternative explanations of the observed relationships with the different analyses, for example with respect to market conditions. In particular, we checked whether there was any correlation between the monthly performance of the Italian stock market and the monthly average of the distances of the stocks under observation. More specifically, we checked for the presence of significant correlations between R_m, the monthly return of the market index (FTSE Italy All-Share Index), and changes in intertextual distances. For each distance considered (Jaccard, Minmax and Delta), the variation is calculated as follows: (distance_month_t – distance_month_t-1)/(distance_month_t-1). We also calculated the correlation between the changes in distances calculated in this way and R_m in month t-1. In no case did we find significant correlations, so market conditions do not seem to relate in any way to the variation in intertextual distances between the articles covered by the column.

Furthermore, the study explores the potential of TD in constructing stock portfolios capable of generating returns beyond those explained by systematic risk factors. Employing pricing models, we assess whether portfolios constructed based on TD exhibit statistically significant excess performance. The findings indicate that the low TD portfolio consistently demonstrates the capacity to generate excess returns, while the high TD portfolio does not, particularly evident in data from January 2015 onwards. Moreover, the study extends its analysis to an out-of-sample context, reaffirming the robustness of TD as a selection criterion for stock portfolios. These results collectively underscore the predictive efficacy of textual similarity in stock selection, encompassing not only earning reports but also newspaper articles.

Declarations of interest: All authors declare that they have no conflicts of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Adosoglou, G., Lombardo, G. & Pardalos, P. M. (2021). Neural Network Embeddings on Corporate Annual Filings for Portfolio Selection. Expert Systems with Applications, 164, 114053. https://doi.org/10.1016/j.eswa.2020.114053.

Albert, Jr., R. & Smaby, T. (1996). Market Response to Analyst Recommendations in the 'Dartboard' Column: The Information and Price-Pressure Effects. Review of Financial Economics, 5, 59–74.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. Literary and Linguistic Computing 23(2), 131–47.

Baker, M. & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. Journal of Finance 61(4), 1645–80.

Barber, B. M. & Loeffler, D. (1993). The 'Dartboard' Column: Second-Hand Information and Price Pressure. Journal of Financial and Quantitative Analysis 28, 273–84.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. (2001). Bad Is Stronger Than Good. Review of General Psychology 5(4), 323–70.

Beneish, M. D. (1991). Stock Prices and the Dissemination of Analysts' Recommendations. Journal of Business 64, 393–416.

Bochkay, K., Brown, S. V., Leone, A. J. & Tucker, J. W. (2023). Textual Analysis in Accounting: What's Next? Contemporary Accounting Research 40(2), 765–805. https://doi.org/10.1111/1911-3846.12825.

Brown, S. V. & Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. Journal of Accounting Research 49(2), 309–46. https://doi.org/10.1111/1911-3846.12825.

Burrows, J. F. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing 17(3), 267–87.

Bybee, L., Kelly, B. & Su, Y. (2023). Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text. The Review of Financial Studies 36(12), 4759–87. https://doi.org/10.1093/rfs/hhad042.

Calomiris, C. W. & Mamaysky, H. (2019). How News and Its Context Drive Risk and Returns around the World. Journal of Financial Economics 133(2), 299–336. https://doi.org/10.1016/j.jfineco.2018.11.009.

Chen, H., De, P., Hu, Y. & Hwang, B. H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. Review of Financial Studies 27(5), 1367–1403. https://doi.org/10.1093/rfs/hhu001.

Cohen, L., Malloy, C. & Nguyen, Q. (2020). Lazy Prices. Journal of Finance 75(3), 1371–1415. https://doi.org/10.1111/jofi.12885.

Costa, L. da F. (2021). Further Generalizations of the Jaccard Index (arXiv:2110.09619). arXiv. http://arxiv.org/abs/2110.09619.

Das, S. R. (2014). Text and Context: Language Analytics in Finance. Foundations and Trends in Finance 8(3). Boston: Now Publishers.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Dyer, T. A., Roulstone, D. T. & Van Buskirk, A. (2024). Disclosure Similarity and Future Stock Return Comovement Management Science 70(7), 4762–80. https://doi.org/10.1287/mnsc.2023.4915.

Dzieliński, M., Wagner, A. & Zeckhauser, R. (2017). Straight Talkers and Vague Talkers: The Effects of Managerial Style in Earnings Conference Calls. NBER Working Papers No. 23425.

Eder, M. (2015). Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from Patrologia Latina. In Digital Humanities 2015: Conference Abstracts. https://zenodo.org/record/1321296#.XzK5APgzagQ.

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. The R Journal 8 (1), 107–121.

Eder, M., Rybicki, J. & Kestemont, M. (2017). 'Stylo': A Package for Stylometric Analyses. Computational Stylistic Group 20.

Enke, B., Graeber, T., Oprea, R. & Yang, J. (2024). Behavioral Attenuation. NBER Working Paper No. 32973, September.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielstrom, S., Schoch, C. & Vitt, T. (2017). Understanding and Explaining Delta Measures for Authorship Attribution. Digital Scholarship in the Humanities 32 (suppl. 2), 4–16.

Fama, E. & French, K. (1993). Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics 33 (1), 3–56.

Farmer, L.E., Schmidt, L. & Timmermann, A. (2023). Pockets of Predictability. Journal of Finance 78 (3), 1279–1341. https://doi.org/10.1111/jofi.13229.

Feldman, R., Govindaraj, S., Livnat, J. & Segal, B. (2010). Management's Tone Change, Post Earnings Announcement Drift and Accruals. Review of Accounting Studies 15, 915–953. https://doi.org/10.1007/s11142-009-9111-x.

García, D. (2013). Sentiment during Recessions. Journal of Finance 68, 1267–1300.

Gentzkow, M., Kelly, B. & Taddy, M. (2019). Text as Data. Journal of Economic Literature 57, 535–574.

Glasserman, P. and Mamaysky, H. (2019). Does Unusual News Forecast Market Stress? Journal of Financial and Quantitative Analysis 54 (5), 1937–1974. https://doi.org/10.1017/S0022109019000127.

Greene, W. H. (1993) *Econometric Analysis*, 2nd ed., Macmillan.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference & Prediction. New York: Springer.

Huang, A., Wu, W. & Yu, T. (2020). Textual Analysis for China's Financial Markets: A Review and Discussion. China Finance Review International 10 (1), 1–15. https://doi.org/10.1108/CFRI-08-2019-0134.

Jaccard, P. (1901). Distribution de la Flore Alpine dans le Bassin des Dranses et dans Quelques Régions Voisines. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241–272.

Jannidis, F., Schöch, C. & Pielstrom, S. (2015). Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures. In Digital Humanities 2015: Conference Abstracts. http://dh2015.org/abstracts.

Juola, P. & Baayen, H. (2005). A Controlled-Corpus Experiment in Authorship Attribution by Cross-Entropy. Literary and Linguistic Computing 20 (1): 59–67.

Ke, Z.T., Kelly, B. & Xiu, D. (2019). Predicting Returns with Text Data. Becker Friedman Institute for Economics at UChicago, WP no. 2019-69.

Kearney, C. & Liu, S. (2014). Textual Sentiment in Finance: A Survey of Methods and Models. International Review of Financial Analysis 33, 171–185. https://doi.org/10.1016/j.irfa.2014.02.006.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. & Daelemans, W. (2016). Authenticating the Writings of Julius Caesar. Expert Systems with Applications 63, 86–96. https://doi.org/10.1016/j.eswa.2016.06.029.

Kogan, S. & Meursault, V. (2021). Corporate Disclosure: Facts or Opinions? Working Papers 21-40, Federal Reserve Bank of Philadelphia. https://doi.org/10.21799/frbp.wp.2021.40.

Koppel, M. & Winter, Y. (2014). Determining If Two Documents Are Written by the Same Author. Journal of the Association for Information Science and Technology 65 (1), 178–187. https://doi.org/10.1002/asi.22954.

Lewis, C. & Young, S. (2019). Fad or Future? Automated Analysis of Financial Text and Its Implications for Corporate Reporting. Accounting and Business Research 49 (5), 587–615. https://doi.org/10.1080/00014788.2019.1611730.

Li, F. (2011). Textual Analysis of Corporate Disclosures: A Survey of the Literature. Journal of Accounting Literature 29, 143–165.

Liu, P., Smith, S. D. & Syed, A. A. (1990). Stock price reactions to theWall Street Journal's securities recommendations, Journal of Financial and Quantitative Analysis 25, 399–410.

Liu, P., Smith, S. D. and Syed, A. A. (1992). The impact of insider trading scandal on the information content of the Wall Street Journal's 'Heard on the Street' column, Journal of Financial Research 15, 181–8.

Liu, S. & Han, J. (2020). Media tone and expected stock returns, International Review of Financial Analysis 70, 101522. https://doi.org/10.1016/j.irfa.2020.101522.

Lloyd-Davies, P. and Canes, M. (1978=. Stock prices and the publication of second-hand information, Journal of Business 51, 43–56.

Loughran, T. & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries & 10-Ks. Journal of Finance 66 (1), 35–65.

Loughran, T. & McDonald, B. (2013). IPO First-Day Returns, Offer Price Revisions, Volatility & Form S-1 Language. Journal of Financial Economics 109 (2). 307–326. https://doi.org/10.1016/j.jfineco.2013.02.017.

Loughran, T. & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research 54, 1187–1230.

Loughran, T. & McDonald, B. (2020). Textual Analysis in Finance. Annual Review of Financial Economics 12 (1), 357–375. https://doi.org/10.1146/annurev-financial-012820-032249.

McMillan, D. G. (2021). Forecasting U.S. Stock Returns. The European Journal of Finance 27 (1-2), 86–109. https://doi.org/10.1080/1351847X.2020.1719175.

Muslu, V., S. Radhakrishnan, K. R. Subramanyam & D. Lim. (2015). Forward-Looking MD&A Disclosures and the Information Environment. Management Science 61, 931–948.

Pettengill, G. N. & J. M. Clark. (2001). Estimating Expected Returns in an Event Study Framework: Evidence from the Dartboard Column. Quarterly Journal of Business and Economics 40, 3–21.

Pound, J. & R. Zeckhouser. (1990). Clearly Heard on the Street: The Effect of Takeover Rumors on Stock Prices. Journal of Business 63, 291–308.

Price, S. M., J. S. Doran, D. R. Peterson & B. A. Bliss. (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. Journal of Banking & Finance 36, 992–1011. https://doi.org/10.1016/j.jbankfin.2011.10.013.

Rapach, D. E. & G. Zhou. (2013). Forecasting Stock Returns. In Handbook of Economic Forecasting, vol. 2, part A, chap. 6, 328–383. https://doi.org/10.1016/B978-0-444-53683-9.00006-2.

Reschenhofer, E., M. Kaur Mangat, C. Zwatz & S. Guzmics. (2020). Evaluation of Current Research on Stock Return Predictability. Journal of Forecasting 39, 334–351. https://doi.org/10.1002/for.2629.

Ružička, M. (1958). Anwendung Mathematisch–Statisticher Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). Biologia, Bratislava 13, 647.

Samuelson, W. & R. Zeckhauser. (1988). Status Quo Bias in Decision Making. Journal of Risk and Uncertainty 1, 7–59.

Sanford, A. (2020). Does Perception Matter in Asset Pricing? Modeling Volatility Jumps Using Twitter-Based Sentiment Indices. Journal of Behavioral Finance 23 (3), 262–280. https://doi.org/10.1080/15427560.2020.1866573.

Silva de Carvalho, R., da Fonseca Tonin, J.M. & Sanches, S.L.R. (2023). Stock Return Predictability Based on Textual Sentiment Analysis: A Review. Revista de Educação e Pesquisa em Contabilidade, Journal of Education and Research in Accounting 17 (2), 140–160. http://dx.doi.org/10.17524/repec.v17i2.3111.

Smith, G. (2018). Step Away from Stepwise. Journal of Big Data 5 (1), 1–12.

Sonkavde, G., Dharrao D.S., Bongale, A.M., Deokate, S.T., Doreswamy, D. & Bhat, S.K. (2023). Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic

Review, Performance Analysis and Discussion of Implication. International Journal of Financial Studies 11 (3), 94. http://doi.org/10.3390/ijfs11030094.

Stalla-Bourdillon, A. (2022). Stock Return Predictability: Comparing Macro- and Micro-Approaches. Banque de France, October, WP #891. https://publications.banque-france.fr/en.

Syed, A. A., Liu, P. and Smith, S. D. (1989). The Exploitation of Inside Information at The Wall Street Journal—A Test of Strong Form Efficiency. Financial Review 24. 567–579.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62 (3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x.

Tetlock, P. C. (2011). All the News That’s Fit to Reprint: Do Investors React to Stale Information? The Review of Financial Studies 24 (5), 1481–1512. https://doi.org/10.1093/rfs/hhq141.

Xing, F.Z., Cambria, E. & Welsch, R.E. (2018). Natural Language-Based Financial Forecasting: A Survey. Artificial Intelligence Review 50 (1), 49–73. https://doi.org/10.1007/s10462-017-9588-9.

Table 1. Cross-section analysis: Descriptive statistics of the variables (January 2005 – June 2022)

	Obs.	Mean	Median	Min	Max	SD
Panel 1: Dependent variables – Buy and Hold Abnormal Returns						
BHAR 3 months	601	0.020	0.005	-0.418	0.748	0.132
BHAR 6 months	601	0.031	0.014	-0.729	1.476	0.207
BHAR 9 months	583	0.044	0.017	-0.889	1.276	0.260
BHAR 12 months	546	0.041	0.005	-0.960	1.518	0.295
Panel 2: Independent variables – Text Distance measures						
Minmax	601	0.771	0.763	0.617	0.928	0.057
Jaccard	601	0.781	0.774	0.620	0.926	0.054
Delta	601	0.147	0.143	0.089	0.274	0.028
Panel 3: Control variables						
Size (ln)	601	7.68	7.72	0.85	11.53	1.39
PBV	601	2.80	1.89	0.17	30.22	2.97
Past performance	601	0.083	0.086	-1.383	0.850	0.243
Beta	601	0.818	0.804	-0.085	2.398	0.363
Days	601	533	392	119	3794	408
New	601	0.131	0.000	0.000	1.000	0.338
AR(0)	601	0.002	0.001	-0.110	0.085	0.019
Text sentiment (TS)	601	0.001	0.000	-0.051	0.055	0.014
Text uncertainty (TU)	601	-0.009	-0.009	-0.024	0.000	0.004
Recommendation	601	0.401	0.000	0.000	1.000	0.491

The table reports descriptive statistics on (i) Buy and Hold Abnormal Returns (Panel 1 – dependent variables, (ii) Text Distance Measures: Delta, Jaccard, and Minmax (Panel 2 – independent variables), (iii) Company size (natural logarithm), Price-to-book value (PBV), Past performance, Company's beta, Days elapsed from the previous article, New columnist, Abnormal return on the day after the publication (AR(0)), Text sentiment, Text Uncertainty, Presence of analysts' Recommendation (Panel 3 – control variables).

Table 2. Abnormal returns for portfolios with Low TD and High TD (January 2005 – June 2022).
In-sample approach.

Text Distance measures	Total obs	LTD (%)	Obs	HTD (%)	Obs	Δ (%)	t-test (*p*)		MW (*p*)		KS (*p*)	
	BHAR(3) - Buy and Hold Abnormal Returns over a period of 3 months											
Minmax	601	2.53	301	1.40	300	1.13	0.294		0.317		0.258	
Jaccard	601	2.77	301	1.17	300	1.60	0.138		0.209		0.183	
Delta	601	1.77	301	2.16	300	-0.39	0.716		0.361		0.158	
	BHAR(6) - Buy and Hold Abnormal Returns over a period of 6 months											
Minmax	601	5.74	301	0.52	300	5.22	0.002	**	0.005	**	0.004	**
Jaccard	601	5.56	301	0.71	300	4.85	0.004	**	0.016	*	0.018	*
Delta	601	4.30	301	1.97	300	2.33	0.168		0.444		0.484	
	BHAR(9) - Buy and Hold Abnormal Returns over a period of 9 months											
Minmax	583	6.44	289	2.33	294	4.12	0.055	°	0.049	*	0.058	
Jaccard	583	6.52	288	2.27	295	4.25	0.048	*	0.048	*	0.074	
Delta	583	5.20	291	3.54	292	1.65	0.443		0.734		0.538	
	BHAR(12) - Buy and Hold Abnormal Returns over a period of 12 months											
Minmax	546	9.09	263	-0.40	283	9.50	0.000	**	0.000	**	0.001	**
Jaccard	546	8.89	263	-0.22	283	9.11	0.000	**	0.001	**	0.001	**
Delta	546	5.37	267	3.02	279	2.36	0.355		0.870		0.911	

LTD = portfolio of stocks with Text Distance below the median value. HTD = portfolio of stocks with Text Distance above the median value. Median over the period January 2005 – June 2022. In-sample approach = portfolios are formed relying on a metric (the median TD 2005–2022) unknown ex-ante, that is when the column is classified as Low or High distance. T-test = t-test for Equality of Means. MW = Mann-Whitney U test. KS = Kolmogorov-Smirnov test. (*p*) = probability value. **, * and ° denote significance at the 1%, 5%, and 10% levels.

Table 3. Abnormal returns for portfolios with Low TD and High TD (January 2007 – June 2022). Out-of-sample approach (rolling-basis median).

Text Distance measures	Total obs	LTD (%)	Obs	HTD (%)	Obs	Δ (%)	t-test (*p*)		MW (*p*)		KS (*p*)	
	BHAR(3) - Buy and Hold Abnormal Returns over a period of 3 months											
Minmax	572	2.01	391	2.05	181	-0.04	0.972		0.954		0.961	
Jaccard	572	2.21	402	1.58	170	0.62	0.641		0.467		0.525	
Delta	572	2.27	368	1.58	204	0.69	0.560		0.842		0.928	
	BHAR(6) - Buy and Hold Abnormal Returns over a period of 6 months											
Minmax	572	4.55	391	0.41	181	4.13	0.028	*	0.057	°	0.070	°
Jaccard	572	4.28	402	0.78	170	3.50	0.079	°	0.105		0.174	
Delta	572	5.27	368	-0.43	204	5.70	0.001	**	0.013	*	0.017	*
	BHAR(9) - Buy and Hold Abnormal Returns over a period of 9 months											
Minmax	554	5.40	378	2.40	176	3.01	0.224		0.176		0.187	
Jaccard	554	4.93	389	3.32	165	1.61	0.546		0.326		0.332	
Delta	554	5.86	358	1.87	196	3.99	0.088	°	0.229		0.192	
	BHAR(12) - Buy and Hold Abnormal Returns over a period of 12 months											
Minmax	519	6.74	349	-0.12	170	6.86	0.015	*	0.028	*	0.023	*
Jaccard	519	6.10	361	0.82	158	5.28	0.074	°	0.086		0.119	
Delta	519	6.78	329	0.54	190	6.24	0.020	*	0.132		0.112	

LTD = portfolio of stocks with Text Distance below the median value. HTD = portfolio of stocks with Text Distance above the median value. The distances from the 2005-06 period are employed to compute the median distance for categorizing the first article of 2007. Following this, the TD of each subsequent article is juxtaposed with the median calculated across all TDs values from January 2005 up to the present moment (rolling-basis median). The approach is ex-ante and out-of-sample, meaning that the classification of an article is determined solely based on information available at the time of decision-making. T-test = t-test for Equality of Means. MW = Mann-Whitney U test. KS = Kolmogorov-Smirnov test. (*p*) = probability value. **, * and ° denote significance at the 1%, 5%, and 10% levels.

<Table 4 about here>

Table 4. Abnormal returns for portfolios with Low TD and High TD (January 2007 – June 2022). Out-of-sample approach (yearly revised median).

Text Distance measures	Total obs	LTD (%)	Obs	HTD (%)	Obs	Δ (%)	t-test (*p*)		MW (*p*)		KS (*p*)	
	BHAR(3) - Buy and Hold Abnormal Returns over a period of 3 months											
Minmax	572	3.08	305	0.82	267	2.26	0.046	*	0.059	°	0.153	
Jaccard	572	3.05	306	0.83	266	2.22	0.051	*	0.034	*	0.024	*
Delta	572	2.29	318	1.68	254	0.61	0.587		0.717		0.339	
	BHAR(6) - Buy and Hold Abnormal Returns over a period of 6 months											
Minmax	572	5.00	305	1.23	267	3.77	0.033	*	0.011	*	0.029	*
Jaccard	572	4.30	306	2.02	266	2.28	0.199		0.075	°	0.088	°
Delta	572	5.07	318	0.95	254	4.12	0.019	*	0.036	*	0.053	*
	BHAR(9) - Buy and Hold Abnormal Returns over a period of 9 months											
Minmax	554	6.39	295	2.23	259	4.17	0.063	°	0.025	*	0.032	*
Jaccard	554	5.87	295	2.83	259	3.05	0.175		0.110		0.220	
Delta	554	6.68	310	1.61	244	5.06	0.025	*	0.013	*	0.013	*
	BHAR(12) - Buy and Hold Abnormal Returns over a period of 12 months											
Minmax	519	8.80	278	-0.48	241	9.28	0.000	**	0.000	**	0.001	**
Jaccard	519	8.62	283	-0.45	236	9.07	0.001	**	0.001	**	0.002	**
Delta	519	8.22	289	-0.19	230	8.41	0.001	**	0.003	**	0.017	*

LTD = portfolio of stocks with Text Distance below the median value. HTD = portfolio of stocks with Text Distance above the median value. The distances from the 2005-06 period are utilized to compute the median distance for classifying all articles published in 2007. From 2008 onward, the TD of each new article is compared with the median of the preceding calendar year, with revisions made on a yearly basis (yearly revised median). The approach is ex-ante and out-of-sample, meaning that the classification of an article is determined solely based on information available at the time of decision-making. T-test = t-test for Equality of Means. MW = Mann-Whitney U test. KS = Kolmogorov-Smirnov test. (*p*) = probability value. **, * and ° denote significance at the 1%, 5%, and 10% levels.

Table 5. Abnormal returns for portfolios with Low TD and High TD (January 2015 – June 2022). Out-of-sample approach (rolling-basis median).

Text Distance measures	Total obs	LTD (%)	Obs	HTD (%)	Obs	Δ (%)	t-test (*p*)		MW (*p*)		KS (*p*)	
	BHAR(3) - Buy and Hold Abnormal Returns over a period of 3 months											
Minmax	281	3.89	115	0.57	166	3.33	0.034	*	0.062	°	0.216	
Jaccard	281	3.22	125	0.89	156	2.33	0.134		0.212		0.215	
Delta	281	1.47	192	2.93	89	-1.46	0.386		0.502		0.676	
	BHAR(6) - Buy and Hold Abnormal Returns over a period of 6 months											
Minmax	281	6.05	115	2.36	166	3.70	0.130		0.041	*	0.119	
Jaccard	281	5.23	125	2.78	156	2.46	0.312		0.152		0.282	
Delta	281	4.45	192	2.62	89	1.83	0.477		0.522		0.526	
	BHAR(9) - Buy and Hold Abnormal Returns over a period of 9 months											
Minmax	265	5.94	106	1.54	159	4.40	0.167		0.053	°	0.124	
Jaccard	265	6.19	114	1.12	151	5.06	0.108		0.032	*	0.065	
Delta	265	3.85	182	2.10	83	1.75	0.627		0.366		0.355	
	BHAR(12) - Buy and Hold Abnormal Returns over a period of 12 months											
Minmax	242	11.86	99	-2.01	143	13.87	0.001	**	0.001	**	0.005	**
Jaccard	242	12.72	108	-3.63	134	16.36	0.000	**	0.000	**	0.000	**
Delta	242	5.12	165	0.55	77	4.56	0.303		0.343		0.644	

LTD = portfolio of stocks with Text Distance below the median value. HTD = portfolio of stocks with Text Distance above the median value. The articles from 2016 are categorized based on the median TD from 2015. Following this, the TD of each new articles is compared to the median calculated from all TD values recorded from January 2005 up to the present moment (rolling-basis median). The approach is ex-ante and out-of-sample, meaning that the classification of an article is determined solely based on information available at the time of decision-making. T-test = t-test for Equality of Means. MW = Mann-Whitney U test. KS = Kolmogorov-Smirnov test. (*p*) = probability value. **, * and ° denote significance at the 1%, 5%, and 10% levels.

Table 6. Abnormal returns for portfolios with Low TD and High TD (January 2015 – June 2022). Out-of-sample approach (yearly revised median).

Text Distance measures	Total obs	LTD (%)	Obs	HTD (%)	Obs	Δ (%)	t-test (*p*)		MW (*p*)		KS (*p*)	
	BHAR(3) - Buy and Hold Abnormal Returns over a period of 3 months											
Minmax	281	3.05	128	0.99	153	2.05	0.187		0.283		0.182	
Jaccard	281	3.34	129	0.73	152	2.60	0.093	°	0.153		0.092	°
Delta	281	1.15	165	3.03	116	-1.88	0.233		0.272		0.237	
	BHAR(6) - Buy and Hold Abnormal Returns over a period of 6 months											
Minmax	281	4.74	128	3.14	153	1.60	0.512		0.178		0.218	
Jaccard	281	4.94	129	2.96	152	1.99	0.415		0.178		0.274	
Delta	281	4.68	165	2.71	116	1.97	0.422		0.623		0.929	
	BHAR(9) - Buy and Hold Abnormal Returns over a period of 9 months											
Minmax	265	4.05	118	2.70	147	1.35	0.666		0.229		0.513	
Jaccard	265	4.85	118	2.06	147	2.79	0.373		0.145		0.258	
Delta	265	4.94	157	0.92	108	4.01	0.218		0.111		0.182	
	BHAR(12) - Buy and Hold Abnormal Returns over a period of 12 months											
Minmax	242	10.07	111	-1.76	131	11.83	0.003	**	0.002	**	0.005	**
Jaccard	242	10.29	112	-2.04	130	12.33	0.002	**	0.001	**	0.008	**
Delta	242	7.24	143	-1.50	99	8.74	0.031	*	0.031	*	0.144	

LTD = portfolio of stocks with Text Distance below the median value. HTD = portfolio of stocks with Text Distance above the median value. The articles from 2016 are categorized based on the median TD from 2015. Subsequently, the TD of each new article is compared to the median value of the preceding calendar year, with annual revisions (yearly revised median). The approach is ex-ante and out-of-sample, meaning that the classification of an article is determined solely based on information available at the time of decision-making. T-test = t-test for Equality of Means. MW = Mann-Whitney U test. KS = Kolmogorov-Smirnov test. (*p*) = probability value. **, * and ° denote significance at the 1%, 5%, and 10% levels.

Table 7. Regressions for Buy and Hold Abnormal Returns: 12-month horizon (January 2005 – June 2022).

	Model (1)	Model (1a)	Model (2)	Model (2a)	Model (3)	Model (3a)
Dependent variable	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)
Regressors:						
Intercept	1.430**	1.240**	1.425**	1.271**	0.763**	0.611**
	(0.305)	(0.236)	(0.319)	(0.250)	(0.190)	(0.165)
TD Minmax	-1.361**	-1.422**				
	(0.462)	(0.326)				
New*TD Minmax	-0.540					
	(1.810)					
TD Jaccard			-1.302**	-1.444**		
			(0.481)	(0.342)		
New*TD Jaccard			-0.727			
			(1.981)			
TD Delta					-0.427	
					(0.759)	
New*TD Delta					-0.227	
					(1.619)	
Size (ln)	-0.016	-0.016°	-0.015	-0.016°	-0.018°	-0.025**
	(0.010)	(0.009)	(0.010)	(0.009)	(0.010)	(0.009)
PBV	-0.001		-0.001		-0.002	
	(0.004)		(0.004)		(0.004)	
Past performance	0.080		0.079		0.101°	0.098°
	(0.054)		(0.054)		(0.054)	(0.052)
Beta	-0.021		-0.022		-0.035	
	(0.038)		(0.038)		(0.038)	
Days (ln)	-0.036		-0.040		-0.078**	-0.063*
	(0.031)		(0.031)		(0.029)	(0.025)
New	0.643	0.155**	0.795	0.147**	0.132	
	(1.572)	(0.051)	(1.728)	(0.051)	(0.304)	
AR(0)	1.089	1.204°	1.131°	1.237°	1.280°	1.317*
	(0.673)	(0.659)	(0.670)	(0.659)	(0.670)	(0.663)
Text sentiment	0.231		0.141		-0.089	
	(1.068)		(1.068)		(1.086)	
Text uncertainty	1.373		1.541		1.730	
	(3.275)		(3.278)		(3.322)	
Recommendation	0.017		0.010		-0.044	
	(0.040)		(0.040)		(0.036)	
R^2 (adjusted)	*0.053*	*0.057*	*0.051*	*0.055*	*0.037*	*0.041*
F	3.523**	9.201**	3.409**	8.906**	2.747**	6.792**
Obs	546	546	546	546	546	546

**, * and ° denote significance at the 1%, 5%, and 10% levels. The table reports six models per the Buy and Hold Abnormal Returns with a 12-months horizon (BHAR(12)) from January 2005: OLS regressions (1,2,3) using the Minmax, or Jaccard, or Delta distances and their interactions with the 'New' dichotomous variable (new journalist) respectively, and all the other control variables; backward selection regressions based on AIC (1a, 2a, 3a) starting from the same set of variables. We present the estimated coefficients, the standard errors, and the significance levels associated with each explanatory variable. The adjusted R-squared, the F-statistics and the number of observations is reported at the bottom of the table.

Table 8. Regressions for Buy and Hold Abnormal Returns: 12-month horizon (January 2015 – June 2022).

	Model (1)	Model (1a)	Model (2)	Model (2a)	Model (3)	Model (3a)
Dependent variable	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)	BHAR(12)
Regressors:						
Intercept	1.730**	1.534**	1.511**	1.436**	0.808**	0.735**
	(0.456)	(0.372)	(0.460)	(0.386)	(0.267)	(0.217)
TD Minmax	-1.764*	-1.868**				
	(0.727)	(0.514)				
New*TD Minmax	-0.480					
	(8.506)					
TD Jaccard			-1.318°	-1.703**		
			(0.722)	(0.526)		
New*TD Jaccard			-1.716			
			(9.421)			
TD Delta					0.312	
					(1.064)	
New*TD Delta					1.373	
					(9.588)	
Size (ln)	-0.017		-0.018		-0.021	
	(0.015)		(0.015)		(0.015)	
PBV	0.005		0.004		0.003	
	(0.006)		(0.006)		(0.006)	
Past performance	0.017		0.025		0.046	
	(0.083)		(0.083)		(0.083)	
Beta	-0.071	-0.120*	-0.076	-0.124*	-0.091	-0.142**
	(0.059)	(0.053)	(0.059)	(0.054)	(0.059)	(0.053)
Days (ln)	-0.031		-0.044		-0.089*	-0.092*
	(0.047)		(0.047)		(0.043)	(0.036)
New	0.549		1.586		-0.292	
	(7.536)		(8.343)		(1.765)	
AR(0)	0.956		0.870		0.874	
	(1.087)		(1.093)		(1.105)	
Text sentiment	-0.075		0.125		0.746	
	(2.060)		(2.068)		(2.067)	
Text uncertainty	1.785		2.487		3.166	
	(4.727)		(4.725)		(4.751)	
Recommendation	-	-	-	-	-	-
	(-)	(-)	(-)	(-)	(-)	(-)
R^2 (adjusted)	*0.059*	*0.075*	*0.051*	*0.066*	*0.039*	*0.053*
F	2.589**	12.25**	2.341**	10.84**	2.018*	8.762**
Obs	277	277	277	277	277	277

**, * and ° denote significance at the 1%, 5%, and 10% levels. The table reports six models per the Buy and Hold Abnormal Returns with a 12-months horizon (BHAR(12)) from January 2015: OLS regressions (1,2,3) using the Minmax, or Jaccard, or Delta distances and their interactions with the 'New' dichotomous variable (new journalist) respectively, and all the other control variables except 'Recommendation' (no more available in this horizon); backward selection regressions based on AIC (1a, 2a, 3a) starting from the same set of variables. We present the estimated coefficients, the standard errors, and the significance levels associated with each explanatory variable. The adjusted R-squared, the F-statistics and the number of observations is reported at the bottom of the table.

Table 9. Monthly extra returns (alpha) for portfolios with Low TD and High TD.
In-sample approach with Minmax Text Distances.

Pricing model	Market Model (4)	Market Model (5)	Market Model (6)	3-factor MM (7)	3-factor MM (8)	3-factor MM (9)
Dependent variable	LTD	HTD	LTD-HTD	LTD	HTD	LTD-HTD
Panel A: January 2005 – June 2022						
Regressors (coefficients):						
Intercept (α_p) %	0.515**	0.130	0.385°	0.376*	0.011	0.365°
	(0.195)	(0.189)	(0.218)	(0.163)	(0.158)	(0.212)
Market return (β_p)	0.884**	0.951**	-0.067°	0.942**	0.987**	-0.046
	(0.032)	(0.031)	(0.035)	(0.027)	(0.026)	(0.036)
VmG return (β_{p2})	---	---	---	-0.088*	0.108**	-0.196**
				(0.041)	(0.040)	(0.053)
SmB return (β_{p3})	---	---	---	0.422**	0.482**	-0.060
				(0.053)	(0.051)	(0.069)
R^2 (adjusted) %	79.77	82.86	1.27	85.87	88.10	6.71
F	777.6**	953.3**	3.5°	400.1**	487.2**	5.7**
Obs	198	198	198	198	198	198
Panel B: January 2015 – June 2022						
Intercept (α_p) %	0.989**	0.155	0.835*	0.763**	-0.023	0.786**
	(0.293)	(0.228)	(0.253)	(0.246)	(0.189)	(0.255)
Market return (β_p)	0.841**	1.010**	-0.168**	0.892**	1.044**	-0.153**
	(0.049)	(0.038)	(0.042)	(0.042)	(0.033)	(0.044)
VmG return (β_{p2})	---	---	---	0.020	0.055	-0.035
				(0.063)	(0.048)	(0.065)
SmB return (β_{p3})	---	---	---	0.556**	0.475**	0.082
				(0.093)	(0.071)	(0.096)
R^2 (adjusted) %	74.40	87.32	12.74	82.23	91.52	12.75
F	294.5*	696.8*	15.8*	156.7*	364.1*	5.9*
Obs	102	102	102	102	102	102

**, * and ° denote significance at the 1%, 5%, and 10% levels. OLS time-series regressions on monthly returns. The table presents alpha and beta coefficients for three portfolios (LTD, HTD, LTD-HTD) under two pricing models. Columns (4) to (6) correspond to the standard Market Model with a single risk factor, the stock market index. Columns (7) to (9) correspond to the 3-factor Market Model (3-factor MM) featuring three risk factors: the stock market index, the Value minus Growth index (VmG), and the Small minus Big index (SmB). We provide the estimated coefficients along with their associated standard errors (in parentheses) and significance levels. The adjusted R^2, F-statistic, significance levels, and the number of observations are presented at the bottom of each regression.

Table 10. Monthly extra returns (alpha) for portfolios with Low TD and High TD. Out-of-sample approach with Minmax Text Distances.

Pricing model	Market Model (10)	Market Model (11)	Market Model (12)	3-factor MM (13)	3-factor MM (14)	3-factor MM (15)
Dependent variable	LTD	HTD	LTD-HTD	LTD	HTD	LTD-HTD
Panel A: January 2007 – June 2022						
Regressors (coefficients):						
Intercept (α_p) %	0.551*	0.178	0.374	0.428*	0.064	0.364
	(0.215)	(0.193)	(0.253)	(0.194)	(0.167)	(0.253)
Market return (β_p)	0.867**	0.953**	-0.086*	0.916**	0.990**	-0.074°
	(0.035)	(0.031)	(0.041)	(0.032)	(0.028)	(0.042)
VmG return (β_{p2})	---	---	---	-0.045	0.080*	-0.126*
				(0.049)	(0.042)	(0.064)
SmB return (β_{p3})	---	---	---	0.397**	0.448**	-0.051
				(0.063)	(0.054)	(0.082)
R^2 *(adjusted) %*	75.64	82.35	1.69	80.35	86.82	2.63
F	612.8**	920.2**	4.4*	269.5**	433.5**	2.77**
Obs	198	198	198	198	198	198
Panel B: January 2015 – June 2022						
Intercept (α_p) %	0.777**	0.211	0.566°	0.554*	0.029	0.525°
	(0.272)	(0.272)	(0.288)	(0.224)	(0.238)	(0.291)
Market return (β_p)	0.876**	1.005**	-0.129**	0.929**	1.039**	-0.110*
	(0.046)	(0.046)	(0.048)	(0.039)	(0.041)	(0.050)
VmG return (β_{p2})	---	---	---	-0.009	0.071	-0.081
				(0.057)	(0.061)	(0.074)
SmB return (β_{p3})	---	---	---	0.521**	0.499**	0.022
				(0.084)	(0.090)	(0.110)
R^2 *(adjusted) %*	78.48	82.82	5.73	85.74	87.04	5.68
F	369.3**	488.1**	7.1**	203.4**	227.1**	3.0*
Obs	102	102	102	102	102	102

**, * and ° denote significance at the 1%, 5%, and 10% levels. OLS time-series regressions on monthly returns. The table presents alpha and beta coefficients for three portfolios (LTD, HTD, LTD-HTD) under two pricing models. Columns (4) to (6) correspond to the standard Market Model with a single risk factor, the stock market index. Columns (7) to (9) correspond to the 3-factor Market Model (3-factor MM) featuring three risk factors: the stock market index, the Value minus Growth index (VmG), and the Small minus Big index (SmB). We provide the estimated coefficients along with their associated standard errors (in parentheses) and significance levels. The adjusted R^2, F-statistic, significance levels, and the number of observations are presented at the bottom of each regression.

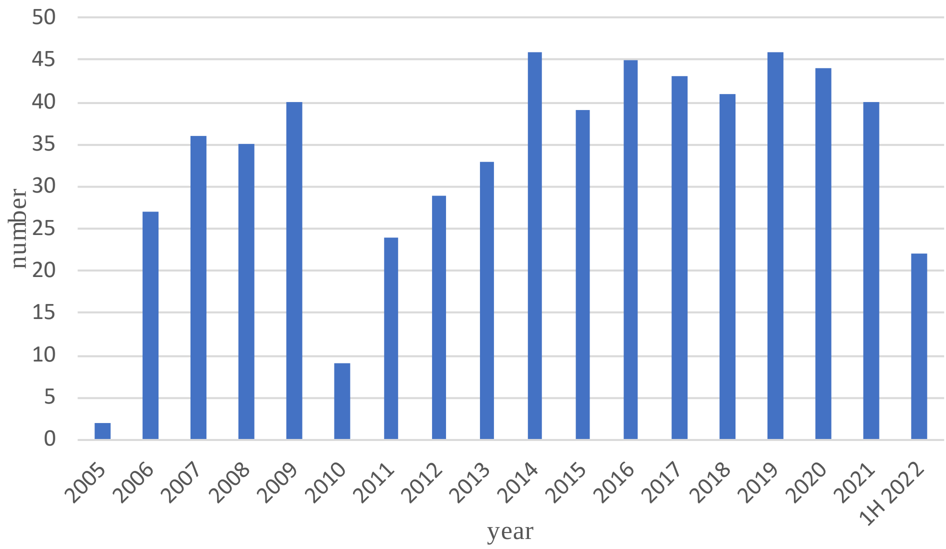

Figure 1 – Number of articles by year.

This figure displays the number of articles published annually in the 'Letter to investors' column for which it was possible to calculate the Text Distance.

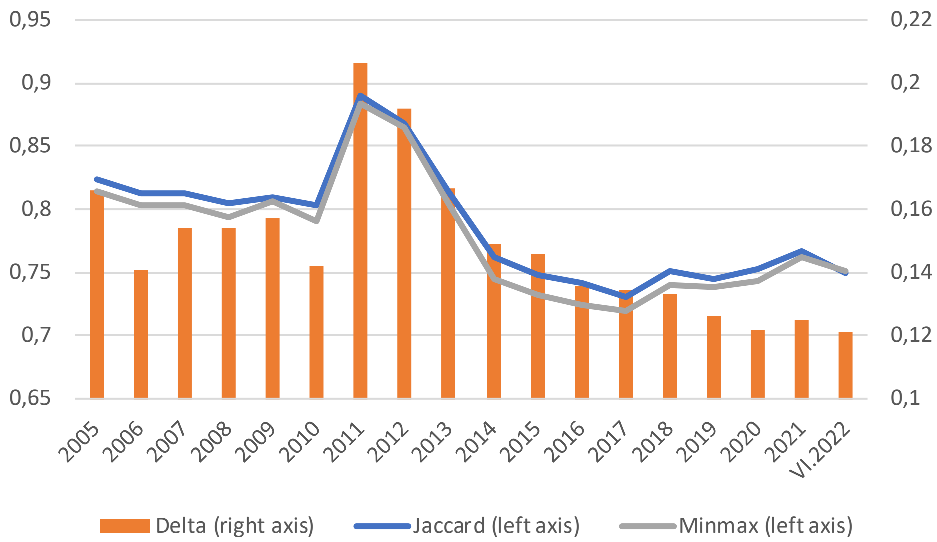

Figure 2 – Text Distances: median values by year.

This figure reports the annual median values of the three Text Distance measures used: Delta, Jaccard, and Minmax.